Pairwise difference estimation of linear panel data

Aquaro, M.

Pairwise difference estimation of linear panel data models

Michele Aquaro

Pairwise difference estimation of linear panel data models

Proefschrift ter verkrijging van de graad van doctor aan Tilburg University, op gezag van de rector magnificus, prof. dr. Ph. Eijlander, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op woensdag 10 april 2013 om 14.15 uur door

Michele Aquaro

geboren op 8 juni 1982 te Martina Franca, Italië.

Promotor:	Prof. dr. Arthur H.O. van Soest
Copromotor:	Dr. Pavel Čížek
Overige leden van de Promotiecommissie:	Prof. dr. Christophe Croux
	Prof. dr. Bertrand Melenberg
	Prof. dr. Tom J. Wansbeek

Ciò che non dai di anima, lo dai di cuore.

(detto popolare)

Contents

Acknowledgments

I would like to express my foremost gratitude to my supervisor Pavel Čížek for his guidance and support. My appreciation goes beyond this thesis and dates back to my early days in Tilburg. Since then, Pavel has been continually active and committed to my research developments and I consider myself to be lucky for having him as supervisor and mentor.

I feel indebted to the committee members Arthur van Soest, Bertrand Melember, Christophe Croux and Tom Wansbeek for their helpful comments and detailed remarks to this thesis. I would also like to take the opportunity to thank Arthur and Bertrand for their advice, encouragement and, ultimately, for my interest in panel data econometrics — interest that originated during an advanced course they were jointly teaching in 2008.

I wish to thank the entire staff at Tilburg University for providing such a friendly and productive work environment. This includes the academic staff, lovely secretaries, CentER graduate officers, the job market committee and the Ph.D. community.

These last years would have not been so enjoyable without Ivana Palinkas, Cristina Majo, Jan Kabatek, Jarda Pasdera, Johannes Binswanger, Marco Della Seta, Miguel Carvalho and Anavi Kinsapla. They all made me very happy and, certainly, a better person.

Thanks to Cristian Bădărînză, Guillaume Schollier and Roberto Romano for their unsweaving support, bright minds and good sense of humor. If I was not shaken by those earthquakes that life sometime throws at us, it was because of their solid friendship. Thanks to Andrea Battinelli and Salvatore Punzi for the enthusiasm they have about my life- and work-developments. Finally, I wish to thank my sister Idria, Peppe and Alessandra for the happy time together.

My parents were not supposed to receive any credit here. They keep me constantly busy over the phone, complaining that I left *Home* to work in countries with rather different weather conditions (my mother) and food traditions (my father), forgetting that *Home* has a double-figure unemployment rate, along with Mediterranean weather and genuine food. But I must admit that I owe them the basic qualities that led to the completion of this work: faith, patience and perseverance. Remaining errors are all mine.

Cambridge, United Kingdom, 27 February 2013.

Chapter 1

Introduction

Panel data sets, also called longitudinal data sets, are sets of data where the same units (for instance individuals, firms, or countries) are observed more than one time. Models that aim at using the specific structure of these data sets are called panel data models. One of the main advantage of using these models is the possibility of including as an explanatory variable some unobserved time-invariant characteristics which are assumed to be heterogeneous across individuals.

More formally, suppose a random sample of n individuals is observed over T time periods, $\{\boldsymbol{z}_{it} = (\boldsymbol{x}_{it}, y_{it})\}_{i=1,t=1}^{n,\ T}$, where $\boldsymbol{x}_{it}$ denotes a $K \times 1$ vector of observable explanatory variables and y_{it} is the dependent variable. Let us consider first the static linear model where the variable y_{it} is modeled as function of $\boldsymbol{x}_{it}$ and some unobservable components η_i and ε_{it}

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \eta_i + \varepsilon_{it} \qquad (t = 1, \cdots, T; \quad i = 1, \cdots, n), \tag{1.1}$$

where $\boldsymbol{\beta}$ is a vector containing the parameters of interest, η_i is an unobservable variable including time invariant individual characteristics of unit i, and ε_{it} is an idiosyncratic shock assumed to be uncorrelated with $\boldsymbol{x}_{it}$. In this thesis, we focus on "fixed-effects" models, i.e. models where no assumptions are made on the distribution of so-called individual effects η_i in (1.1).

Model (1.1) can be extended by allowing y_{it} to depend also on its past values. Using the same notation as before, a linear dynamic panel data models with P lags reads as follows

$$y_{it} = \sum_{p=1}^{P} \alpha_p y_{it-p} + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \eta_i + \varepsilon_{it} \qquad (t = 1, \cdots, T; \quad i = 1, \cdots, n). \tag{1.2}$$

The effects η_i in model (1.2) are assumed to be uncorrelated with the idiosyncratic shocks $\varepsilon_{i1}, \cdots, \varepsilon_{iT}$ for all $i = 1, \cdots, n$. This is the only crucial assumption made on the distribution of η_i in model (1.2). Somehow misleadingly, the majority of the literature has called "fixed-effects' models also the dynamic models (1.2) with the zero correlation assumption of the effects and the idiosyncratic errors (among others see Ahn and Schmidt, 1997).

The related literature for models like (1.1) and (1.2) is vast. For a general historical account of panel data methods in years 1861–1997, see Nerlove (2005, Chapter 1), whereas more recent reviews can be found in Baltagi (2005) Mátyás and Sevestre (2008), and Wooldridge (2001), among others. Unfortunately, almost all literature focuses on models assuming that data are free of outlying or aberrant observations. This is often not the case in reality. The majority of the regression methods used in linear panel data models are very sensitive to data contamination and outliers. This sensitivity can be characterized by various measures of robustness such as the breakdown point, which measures the smallest contaminated fraction of a sample that can arbitrarily change the estimates (Genton and Lucas, 2003; Davies and Gather, 2005). Because the breakdown point of linear estimators is asymptotically zero, many authors stressed the importance of robust and positive breakdown-point methods (e.g., Hampel et al., 1986; Simpson et al., 1992; Ronchetti and Trojani, 2001; Gervini and Yohai, 2002; Wagenvoort and Waldmann, 2002; Maronna et al., 2006; Čížek, 2008). This is even more important in the case of panel data, which can contain individuals with erroneous observations that are masked by the complex structure of the data.

Despite its relevance, the study of robust techniques for panel data seems to be rather limited. The works of Wagenvoort and Waldmann (2002) and Lucas et al. (2007) concentrate on the bounded-influence estimation of static and dynamic panel-data models, respectively. Along with related quantile-regression estimation by Koenker (2004), these methods are generally locally robust, that is, their breakdown point can be arbitrarily close to zero for some kinds of data contamination. The positive breakdown-point methods were proposed only by Bramati and Croux (2007) and Dhaene and Zhu (2009), where the first concentrates on the static panel models and the latter on the dynamic panel models.

The thesis consists of two parts. In the first part, some alternative robust estimation methods are proposed for models (1.1) and (1.2) (Chapter 2 and Chapter 3, respectively). The second part (Chapter 4) is a contribution to the theory of estimation of dynamic models (1.2) when data are assumed not to be contaminated. As we will see later on (Chapter 4), dynamic panel data models can be difficult to estimate even for outlier free data when the stationarity assumption on y_{it} does not hold. The common theme underling the next three chapters is the pairwise difference data transformation, that is a generalization of the first difference operator used to filter out the unobserved effects η_i from Equations (1.1) and (1.2).

1.1 Summary

The thesis is based on the following research papers:

- *Chapter 2:* Aquaro M., and P. Čížek (2013), One-step robust estimation of fixed-effects panel data models, *Computational Statistics & Data Analysis*, 57:1, 536–548.
- *Chapter 3:* Aquaro M., and P. Čížek (2012), Robust estimation of dynamic fixed-effects panel data models, *Working Paper*.
- *Chapter 4:* Aquaro M., and P. Čížek (2012), Pairwise difference estimation of dynamic panel data models, *Working Paper*.

In Chapter 2, a new estimation approach for fixed-effects static panel data models based on two different data transformations is proposed. Considering several robust estimation methods applied to the transformed data, the robust and asymptotic properties of the proposed estimators are derived, including their breakdown points and asymptotic distributions. The finite-sample performance of the existing and proposed methods is compared by means of Monte Carlo simulations.

Chapter 3 extends an existing outlier-robust estimator of linear dynamic panel data models with fixed effects, which is based on the median ratio of two consecutive pairs of first-differenced data. To improve its precision and robust properties, a general procedure based on many pairwise differences and their ratios is designed. The asymptotic distribution of this class of estimators is derived. Further, the breakdown point properties are obtained under contamination by independent additive outliers and the patches of additive outliers and are used to select the pairwise differences that do not compromise the robust properties of the procedure. The proposed estimator is additionally compared with existing methods by means of Monte Carlo simulations.

In Chapter 4, a new estimation procedure of dynamic panel data models with fixed effects is proposed. To improve upon existing estimators, we propose to apply the pairwise-difference data transformation to the generalized method of moments based estimators. A particular focus is given to the long difference (LD) estimation procedure of Hahn et al. (2007), which was proved to retain strong moment conditions even when data are persistent without imposing the covariates to have constant correlation with the effects. The bias and asymptotic distribution of the original LD estimator and its proposed extensions are derived. A simulation study is conducted to assess the finite-sample properties of the estimators.

Chapter 2

One-step robust estimation of fixed-effects panel data models[1]

2.1 Introduction

The panel-data regression models are increasingly popular in applications because each individual cross-sectional unit is observed over time, and consequently, the individual-specific heterogeneity can be accounted for. The majority of the regression methods used in linear panel-data models are based on linear estimators such as least squares (LS), and consequently, are very sensitive to data contamination and outliers. This sensitivity can be characterized by various measures of robustness such as the breakdown point, which measures the smallest contaminated fraction of a sample that can arbitrarily change the estimates (Genton and Lucas, 2003; Davies and Gather, 2005). Because the breakdown point of the linear estimators such as LS is asymptotically zero, many authors stressed the importance of robust and positive breakdown-point methods (e.g., Hampel et al., 1986; Simpson et al., 1992; Ronchetti and Trojani, 2001; Gervini and Yohai, 2002; Wagenvoort and Waldmann, 2002; Maronna et al., 2006; Čížek, 2008). This is even more important in the case of panel data, which can contain individuals with erroneous observations that are masked by the complex structure of the data.

Despite its relevance, the study of robust techniques for panel data seems to be rather limited. The works of Wagenvoort and Waldmann (2002) and Lucas et al. (2007) concentrate on the bounded-influence estimation of static

[1]This chapter is based on Aquaro M., and P. Čížek (2013), One-step robust estimation of fixed-effects panel data models, *Computational Statistics & Data Analysis*, 57:1, 536–548.

and dynamic panel-data models, respectively. Along with related quantile-regression estimation by Koenker (2004), these methods are generally locally robust, that is, their breakdown point can be arbitrarily close to zero for some kinds of data contamination. The positive breakdown-point methods were proposed only by Bramati and Croux (2007) and Dhaene and Zhu (2009), where the first concentrates on the static panel models and the latter on the dynamic panel models. Being interested in the static panel-data models here, Dhaene and Zhu (2009) aiming at dynamic models is not suitable, especially since it strictly relies on additional distributional assumptions (e.g., errors being normal and independent and identically distributed), which rule out serial correlation of errors. On the other hand, the methods proposed by Bramati and Croux (2007) either are not equivariant with respect to various linear data transformations or have to explicitly estimate the fixed effects, causing bias due to the nonlinearity of the procedure if the number of periods is fixed (see Sections 2.2.2 and 2.4 for details). In both cases, the methods are consistent only if the number of time periods increases to infinity, which makes them unsuitable for short panels.

In this paper, we propose an alternative robust estimation approach for linear fixed-effect panel-data models that is equivariant with respect to standard data transformations, that is consistent for data observed only a (small) fixed number of time periods, and that, besides the standard identification assumptions, does not require any particular distributional assumptions (with the exception of the errors having a unimodal distribution). To achieve this, we employ two different data transformations and show that it is possible to apply standard robust estimators of linear regression to the transformed data. Because of the data transformations, the equivariance, robust, and asymptotic properties of the proposed estimators have to be established. All methods are shown to have a positive breakdown point greater or equal to or at least converging to 1/4 and to have asymptotically a normal distribution. Additionally, some of the proposed methods are asymptotically efficient for normal data. At the same time, Monte Carlo experiments indicate that the finite-sample performance of the proposed methods matches the standard within-group LS estimator and the robust properties thus do not adversely affect the precision of estimation using data free of outliers.

The paper is organized as follows. After a survey of the existing fixed-effect panel-data estimators in Section 2.2, two data transformations and the corresponding robust estimators are proposed in Section 2.3, where their robust and asymptotic properties are also examined. The finite-sample properties are studied in Section 2.4 and an empirical example is provided in Section 2.5. The proofs are given in the Appendix.

2.2 Panel data models

In this section, a brief account of some classical panel-data estimators is offered (Section 2.2.1), followed by the discussion of existing robust methods suitable for panel data (Section 2.2.2).

2.2.1 The fixed-effects model

A static linear fixed-effects panel-data model can be described by

$$y_{it} = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}, \qquad i = 1, \ldots, n, \qquad t = 1, \ldots, T, \tag{2.1}$$

where y_{it} denotes the dependent variable, $\boldsymbol{x}_{it} \in \mathbb{R}^p$ contains observable covariates, and the vector $\boldsymbol{\beta} \in \mathbb{R}^p$ represents the parameters of interest. The subscript i could refer to individuals, households, firms, or countries, whereas t indicates the periodicity. The unobservable terms consist of an unobservable individual-specific effect α_i and of the error term ε_{it}, which is assumed to have a zero mean, $\mathrm{E}\left(\varepsilon_{it}|\alpha_i, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}\right) = 0$, and to be independent across individuals; see Wooldridge (2001).

Without additional assumptions about the individial effects α_i and given a fixed number of observed time periods T, the estimation of $\boldsymbol{\beta}$ is straightforward only if α_i's are eliminated from the model equation. The standard within-group transformation rules out the fixed effects by computing the time averages for each individual,

$$\bar{y}_{i\cdot} = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \qquad \bar{\boldsymbol{x}}_{i\cdot} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_{it}, \tag{2.2}$$

and then subtracting them from the original values: $\tilde{y}_{it} = y_{it} - \bar{y}_{i\cdot}$ and $\tilde{\boldsymbol{x}}_{it} = \boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_{i\cdot}$. Model (2.1) then implies the linear relationship $\tilde{y}_{it} = \tilde{\boldsymbol{x}}_{it}^{\top}\boldsymbol{\beta} + \tilde{\varepsilon}_{it}$, which permits estimating the parameter vector $\boldsymbol{\beta}$ by the LS estimate $\hat{\boldsymbol{\beta}}_{nT}^{(\mathrm{LS,mean})}$. The within-group LS estimator is linear, which implies that it is equivariant with respect to scale, regression, and affine transformations: denoting the estimator explicitly as a function of data $\mathcal{T}^{\mathrm{LS}}(\{\boldsymbol{x}_{it}, y_{it}\})$, the scale, regression, and affine equivariance mean that

$\mathcal{T}^{\mathrm{LS}}(\{\boldsymbol{x}_{it}, cy_{it}\}) = c\mathcal{T}^{\mathrm{LS}}(\{\boldsymbol{x}_{it}, y_{it}\})$, $\mathcal{T}^{\mathrm{LS}}(\{\boldsymbol{x}_{it}, y_{it} + \boldsymbol{x}_{it}^{\top}\boldsymbol{v}\}) = \mathcal{T}^{\mathrm{LS}}(\{\boldsymbol{x}_{it}, y_{it}\}) + \boldsymbol{v}$, and $\mathcal{T}^{\mathrm{LS}}(\{\boldsymbol{A}^{\top}\boldsymbol{x}_{it}, y_{it}\}) = \boldsymbol{A}^{-1}\mathcal{T}^{\mathrm{LS}}(\{\boldsymbol{x}_{it}, y_{it}\})$, respectively, for any $c \in \mathbb{R}, \boldsymbol{v} \in \mathbb{R}^p$, and non-singular $\boldsymbol{A} \in \mathbb{R}^{p\times p}$.

Unfortunately, the within-group LS estimator is very sensitive to erroneous observations and outliers as any linear regression LS method. To document this, let us introduce one of the global measures of robustness – the breakdown point. Informally, an estimator is said to break down when the procedure no longer conveys useful information on the data-generating mechanism (Genton and Lucas, 2003). In linear regression models, this general statement is equivalent to saying that the estimates can increase above any bound in the presence of data contamination: for a random sample $Z = \{\boldsymbol{x}_{it}, y_{it}\}_{i=1,t=1}^{n,\ T}$ and an estimator $\mathcal{T}$ of the regression parameters, the finite-sample breakdown point of $\mathcal{T}$ at the sample Z can be defined by (Rousseeuw and Leroy, 1987)

$$\varepsilon_{nT}^{*}(\mathcal{T}; Z) = \frac{1}{nT}\max_{m\geq 0}\left\{m \mid \sup_{Z_m}\|\mathcal{T}(Z) - \mathcal{T}(Z_m)\| < \infty\right\}, \tag{2.3}$$

where the supremum is taken over all choices of Z_m consisting of $(nT - m)$ points from Z and m arbitrary points. The asymptotic breakdown point of $\mathcal{T}$ can be defined as the limit $\varepsilon^*(\mathcal{T}) = \lim_{nT\to\infty}\varepsilon_{nT}^{*}(\mathcal{T}; Z)$, provided that this sample-independent limit exists. It can be at most $1/2$ for regression equivariant estimators (cf. Davies and Gather, 2005). For the within-group LS estimator, the finite-sample breakdown point however does not exceed $1/nT$ and it converges to zero asymptotically.

2.2.2 Robust estimators for panel data

To the best of our knowledge, there are very few studies proposing robust estimators for panel data. Considering the globally robust estimators (i.e., having a positive breakdown point), the two existing contributions are Dhaene and Zhu (2009) and Bramati and Croux (2007). The first one proposes median-based estimators for dynamic fixed-effects models, which strictly require errors being independent and identically distributed across all individuals and time periods and does not allow for heteroscedasticity and serial autocorrelation often encountered in static panel-data models. Thus, the only proposal generally applicable in static fixed-effect panel-data models stems from Bramati and Croux (2007), who adapt two existing high-breakdown point procedures and reach asymptotically a positive breakdown

Table 2.1: The mean squared errors of the within-group LS and WGM estimates based on the mean and median transformations.

M		0		1		10	
# parameters		1	5	1	5	1	5
mean	LS	0.001	0.022	0.001	0.021	0.001	0.021
med	LS	0.001	0.017	0.003	0.065	0.007	4.272
	WGM	0.001	0.005	0.146	0.809	0.046	51.88

1/4: the within-group generalized M-estimator (WGM) and the within-group MS estimator (WMS).

The WMS estimator explicitly estimates the fixed effects, treating them as regression coefficients, and applies the robust MS-estimator of Maronna and Yohai (2000), which is able to deal with a large amount of discrete explanatory variables. This estimator can achieve the breakdown point up to 1/4, is consistent if the number of time periods $T \to \infty$, but there are no results concerning its consistency if the number of periods T is small.

The WGM estimator applies robust estimators to centered data, where Bramati and Croux (2007) replace the non-robust mean in (2.2) by the median. Variables are thus centered using the within-group medians:

$$\tilde{y}_{it} = y_{it} - \mathrm{med}_t\, y_{it}, \qquad \tilde{\boldsymbol{x}}_{it} = \boldsymbol{x}_{it} - \mathrm{med}_t\, \boldsymbol{x}_{it}. \tag{2.4}$$

After centering, a natural approach is to regress $\tilde{y}_{it}$ on $\tilde{\boldsymbol{x}}_{it}$ using a robust regression estimator. Bramati and Croux (2007) suggest to use first the least trimmed squares (LTS) estimator (Rousseeuw, 1984). Given that LTS with the maximum breakdown point has a rather low relative efficiency of 8% for normal data, the reweighted LS strategy is adopted in the second step, where weights are designed so that the breakdown point of the initial LTS estimator is preserved (Rousseeuw and Leroy, 1987).

The complete WGM procedure can asymptotically achieve the breakdown point 1/4. On the other hand, WGM is neither regression nor affine equivariant and its consistency and asymptotic distribution (even for $T \to \infty$) have not been studied yet. This lack of equivariance comes from the nonlinearity of the median transformation. Loosely speaking, equivariance is a desirable property: transforming of data (by scaling or by linear combinations) by a certain factor results in estimates which are transformed by the same factor. On the contrary, if an estimator is not equivariant, transformation of data would change the estimation results. Let us now illustrate how the lack of equivariance complicates the use of WGM in applications for a small T. Consider the following linear panel-data model ($i = 1, \ldots, 100$; $t = 1, 2, 3$)

$$y_{it} = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}, \tag{2.5}$$

where $\boldsymbol{x}_{it} \sim N(0,1)$, $\varepsilon_{it} \sim N(0,1)$, $\alpha_i \sim U(0,10)$, and $\boldsymbol{\beta} = -M \in \mathbb{R}$ or $\boldsymbol{\beta} = (-M, 0, M, 0, -M)^{\top} \in \mathbb{R}^5$. Simulating the data 1000 times and estimating the model for $M = 0, 1$, and 10 by LS and WGM results in the mean squared errors in Table 2.1. Obviously, various levels of the multiplier M do not have any impact on the precision of the within-group LS estimator. Using LS and WGM after removing the individual effects by the median centering however leads to completely different results: the mean squared errors are substantially increasing with the magnitude of the regression coefficients, especially for the model with 5 variables.

2.3 New robust estimators for panel data

To provide panel-data estimators that are equivariant, consistent, and asymptotically normal even if the number of time periods T is small and fixed, we now propose alternative robust estimators of $\boldsymbol{\beta}$ in (2.1) that do not rely on estimating the fixed effects (i.e., the expected values of individual effects). This will be done in two steps. First, the elimination of the unobserved individual effects will be addressed by considering other data transformations than the mean or median centering (Section 2.3.1). Second, in the light of recent contributions in robust statistical theory, the initial LTS estimation (Section 2.3.2) will be followed by new robust and efficient estimators adapted to the panel data setting (Section 2.3.3).

2.3.1 Data transformations

To construct an alternative to the methods discussed in Section 2.2.2, we focus on the first-difference and pairwise-difference transformations instead. The first-difference transformation is already well known in the literature (Wooldridge, 2001). Denoting the first-difference operator by Δ, the model (2.1) can be transformed to

$$\Delta y_{it} = y_{it} - y_{it-1} = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \varepsilon_{it} - \boldsymbol{x}_{it-1}^{\top}\boldsymbol{\beta} - \varepsilon_{it-1} = \Delta\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \Delta\varepsilon_{it}, \tag{2.6}$$

where $i = 1, \ldots, n$ and $t = 2, \ldots, T$ and where no fixed effects α_i appear. Under the strict-exogeneity assumption of model (2.1), $\boldsymbol{\beta}$ is consistently estimated by LS applied to (2.6). This alternative to the within-group estimator, which is the best linear unbiased estimator when error terms ε_{it} are uncorrelated, is preferable only if error terms ε_{it} exhibit a strong positive serial correlation (see Wooldridge, 2001, for details).

Alternatively, one could aim for more accurate estimates than from (2.6) by eliminating individual effects by taking all pairwise differences within each individual. Inspired by Stromberg et al. (2000) and Honoré and Powell (2005), we propose to transform data using the pairwise-difference transformation as $\Delta^s z_{it} = z_{it} - z_{it-s}$, where $s = 1, \ldots, t-1$, for any $t \in \{2, \ldots, T\}$ and $i \in \{1, \ldots, n\}$. Applied to model (2.1), the pairwise-difference transformation yields

$$\Delta^s y_{it} = y_{it} - y_{it-s} = (\boldsymbol{x}_{it} - \boldsymbol{x}_{it-s})^\top \boldsymbol{\beta} + \varepsilon_{it} - \varepsilon_{it-s} = \Delta^s \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + \Delta^s \varepsilon_{it}, \quad (2.7)$$

which also removes the individual-specific variable α_i, but generates a larger sample size $nT(T-1)/2$ instead of $n(T-1)$ in (2.6) since differences for all $s = 1, \ldots, t-1$ are considered.

To handle all transformations in a unified way, let us now introduce a more general notation. Given the original data set $\{\boldsymbol{x}_{it}, y_{it}\}_{i=1,t=1}^{n,\ T}$, let $\{\tilde{\boldsymbol{x}}_{it}, \tilde{y}_{it}\}_{i=1,t=1}^{n,\ T^{(\mathfrak{T})}}$ be the data set created by one of the considered data transformations $\mathfrak{T}$, $\mathfrak{T} \in \{\text{med}, 1\Delta, P\Delta\}$, where med, 1Δ, and $P\Delta$ are shorthand symbols for the median-centering, first-difference, and pairwise-difference transformation and $T^{(\mathfrak{T})} = T, T-1$, and $T(T-1)/2$, respectively.

2.3.2 Initial robust estimator

Once the individual effects have been eliminated, it is of interest to find a proper robust estimator for $\boldsymbol{\beta}$ in (2.1). Similarly to Bramati and Croux (2007), we use initially the LTS estimator, which may be generally defined for the $\mathfrak{T}$-transformed data as

$$\hat{\boldsymbol{\beta}}_{nT}^{(\text{LTS},\mathfrak{T},h_{nT})} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{j=1}^{h_{nT}} r_{(j)}^{2,(\mathfrak{T})}(\boldsymbol{\beta}), \quad (2.8)$$

where $r_{(j)}^{2,(\mathfrak{T})}(\boldsymbol{\beta})$ is the jth smallest order statistics of the squared residuals, the (i,t)th residual equals $r_{it}^{(\mathfrak{T})}(\boldsymbol{\beta}) = \tilde{y}_{it} - \tilde{\boldsymbol{x}}_{it}^\top \boldsymbol{\beta}$, and h_{nT} is the trimming constant, $nT^{(\mathfrak{T})}/2 < h_{nT} \leq nT^{(\mathfrak{T})}$. We assume that the trimming constant, which determines the number $nT^{(\mathfrak{T})} - h_{nT}$ of observations excluded from the objective function, is defined so that $h_{nT}/nT^{(\mathfrak{T})} \to \lambda \in [1/2, 1]$, and thus asymptotically, the $1-\lambda$ fraction of observations is eliminated from the objective function (2.8). The maximum breakdown point is attained for $h_{nT} = \left[nT^{(\mathfrak{T})}/2\right] + [(p+1)/2] + 1$ (Rousseeuw and Leroy, 1987) with $h_{nT}/nT^{(\mathfrak{T})} \to 1/2$, where $[x]$ denotes the integer part of x.

Contrary to the median centering, both the first-difference and pairwise-difference transformations are linear transformations of the data. Therefore, the LTS estimator applied to such transformed data does not lose its equivariance properties contrary to LTS applied to the median-transformed data in Bramati and Croux (2007).

Lemma 1. *If $\mathfrak{T} \in \{1\Delta, P\Delta\}$, then the LTS estimator $\hat{\boldsymbol{\beta}}_{nT}^{(\mathrm{LTS},\mathfrak{T},h_{nT})}$ defined in (2.8) is scale, affine, and regression equivariant.*

Further, the breakdown properties of the LTS estimator in general depend on the employed data transformation, but are similar to or better than the breakdown point of LTS under the median transformation of Bramati and Croux (2007).

Theorem 1. *Let $Z_{nT} = \{\boldsymbol{x}_{it}, y_{it}\}_{i=1,t=1}^{n,\ \ T}$ be a random sample generated according to model (2.1). Further, the transformed data $\{\tilde{\boldsymbol{x}}_{it}, \tilde{y}_{it}\}_{i=1,t=1}^{n,\ \ T^{(\mathfrak{T})}}$ are assumed to be in a general position for $nT^{(\mathfrak{T})} > 3(p+1)$ almost surely, that is, any $p+1$ data points do not lie on the same hyperplane almost surely. Finally, let $\hat{\boldsymbol{\beta}}_{nT}^{(\mathrm{LTS},\mathfrak{T},h_{nT})}$ be the LTS estimator defined in (2.8) for $h_{nT}/(nT^{(\mathfrak{T})}) \to \lambda$ as $nT^{(\mathfrak{T})} \to \infty$. If $h_{nT} \geq \underline{h}_{nT}^{T^{(\mathfrak{T})}} = [nT^{(\mathfrak{T})}/2] + [(p+1)/2] + 1$, then it holds that*

$$\varepsilon_{nT}^{*}\left(\hat{\boldsymbol{\beta}}_{nT}^{(\mathrm{LTS},\mathfrak{T},h_{nT})}; Z_{nT}\right) \geq \frac{nT^{(\mathfrak{T})} - h_{nT}}{2nT^{(\mathfrak{T})}} \cdot \kappa^{(\mathfrak{T})}(T), \tag{2.9}$$

where $\kappa^{(1\Delta)} = [2(T-1)]/[\min\{2, T-1\}T]$ and $\kappa^{(P\Delta)} = 1$. The bound on the breakdown point of LTS tends to $\kappa^{(\mathfrak{T})}(T)(1-\lambda)/2$ for $n \to \infty$, and in particular, to $\kappa^{(\mathfrak{T})}(T)/4$ for $h_{nT} = \underline{h}_{nT}^{T^{(\mathfrak{T})}}$.

In the case of the pairwise difference transformation, it also holds for $h_{nT} \geq \underline{h}_{nT}^{T^{(\mathfrak{T})}}$ that

$$\varepsilon_{nT}^{*}\left(\hat{\boldsymbol{\beta}}_{nT}^{(\mathrm{LTS},P\Delta,h_{nT})}; Z_{nT}\right) \geq \frac{1}{T}\left\lfloor T - \frac{1}{2} - \sqrt{\frac{2h_{nT}}{n} + \frac{1}{4}}\right\rfloor, \tag{2.10}$$

where $2h_{nT}/n \to 2\lambda T^{(\mathfrak{T})} = \lambda T(T-1)$ for $n \to \infty$.

From the breakdown point of view, both proposed data transformations are equivalent for $T = 2$ and the lower bounds (2.9) are equivalent for $T \to \infty$ as they yield the same maximum breakdown point 1/4 analogously to Bramati and Croux (2007). Whereas the pairwise differencing reaches this lower bound breakdown point for any number of time periods T, the first differencing has a smaller breakdown point equal to $(T-1)/(4T)$ for $T \geq 3$. For the pairwise difference estimator, one more lower bound holds and is given in (2.10). This bound provides a good approximation of the breakdown point only if the number of outliers is a multiple of n and can be thus meaningfully used only for $T > 3$. Allowing $T \to \infty$, the breakdown-point bound (2.10) converges to $1 - \sqrt{\lambda}$. This implies that the breakdown of LTS applied to the pairwise differenced data converges to $1 - \sqrt{1/2} \approx 0.29$ for $h_{nT} = \underline{h}_{nT}^{T(\mathfrak{T})}$ and $\lambda = 1/2$, which is higher than the one achieved by other discussed estimators. The second breakdown bound (2.10) can however be further improved by adapting the result of Stromberg et al. (2000, Theorem 3) to panel data, who allowed $\lambda < 1/2$ and proved that the breakdown point of the least trimmed differences estimator equals asymptotically $\min\{\sqrt{\lambda}, 1 - \sqrt{\lambda}\}$. For $T \to \infty$, we could then conjecture that the breakdown point of the pairwise difference LTS estimator can reach 1/2 if $\lambda = 1/4$.

2.3.3 Robust and efficient estimation

Since the LTS estimator with the maximum breakdown point achieves only 8% relative efficiency for normally distributed data, one-step estimators are often employed to improve the precision of estimation without substantially affecting the robust properties of estimation (see also Section 2.2.2). Instead of the simple reweighted LS employed by Bramati and Croux (2007), we propose to use efficient data-adaptive robust methods of Gervini and Yohai (2002) and Čížek (2010).

To introduce the efficient one-step methods, suppose we have the transformed data $\{\tilde{\boldsymbol{x}}_{it}, \tilde{y}_{it}\}$ obtained by transformation $\mathfrak{T} \in \{\mathrm{med}, 1\Delta, P\Delta\}$ and a pair of initial robust estimators of the regression parameters $\hat{\boldsymbol{\beta}}_{nT}^{0}$ and residual scale $\hat{\sigma}_{nT}^{0}$ (e.g., the median absolute deviation). A classical example of a one-step augmentation procedure is the reweighted LS estimator proposed by Rousseeuw and Leroy (1987), which removes the observations having large absolute residuals according to some initial robust fit and then applies LS. Denoting the initial residuals $r_{it}^{(\mathfrak{T})}(\hat{\boldsymbol{\beta}}_{nT}^{0}) = \tilde{y}_{it} - \tilde{\boldsymbol{x}}_{it}^{\top}\hat{\boldsymbol{\beta}}_{nT}^{0}$, the weights determining the inclusion or removal of observations can be defined by

$$\hat{w}_{it}\left(\hat{\boldsymbol{\beta}}^0_{nT}, \hat{\sigma}^0_{nT}; v\right) = I\left(|r^{(\mathfrak{T})}_{it}(\hat{\boldsymbol{\beta}}^0_{nT})/\hat{\sigma}^0_{nT}| < v\right) \tag{2.11}$$

for a constant $v > 0$ (e.g., Gervini and Yohai (2002) suggest $v = 2.5$).

A data-adaptive version of weights (2.11) designed to achieve efficiency for normally distributed data, the robust and efficient weighted least squares (REWLS) estimator, has been proposed by Gervini and Yohai (2002). A data-dependent cut-off point $\hat{v}_{nT}$ to define weights (2.11) is now determined by comparing two distribution functions, F^+ and F^+_0, where the former relates to the standardized absolute residuals $|r^{(\mathfrak{T})}_{it}(\hat{\boldsymbol{\beta}}^0_{nT})/\hat{\sigma}^0_{nT}|$ and the latter is the distribution function assumed for these standardized absolute residuals in the model (2.1). Since F^+ is usually unknown, it is estimated by the empirical distribution function F^+_{nT} of $|r^{(\mathfrak{T})}_{it}(\hat{\boldsymbol{\beta}}^0_{nT})/\hat{\sigma}^0_{nT}|$. The maximum discrepancy $\hat{d}_{nT}$ between F^+_{nT} and F^+_0 in the tail of the distributions can be then measured by

$$\hat{d}_{nT} = \sup_{v\geq\eta}\left\{\left[F^+_0(v) - F^+_{nT}(v)\right]\cdot I\left(F^+_0(v) - F^+_{nT}(v) \geq 0\right)\right\}, \tag{2.12}$$

where η is a large quantile of F^+_0, for example, $\eta = 2.5$ for Gaussian errors with $F_0 \equiv N(0,1)$ (see Gervini and Yohai, 2002). The cutoff point $\hat{v}_{nT}$ is then defined as the $(1-\hat{d}_{nT})$th quantile of the distribution F^+_{nT}: $\hat{v}_{nT} = \min\left\{v \mid F^+_{nT}(v) \geq 1 - \hat{d}_{nT}\right\}$. Finally, the REWLS estimator is obtained using weights (2.11) with $v = \hat{v}_{nT} \geq \eta$:

$$\hat{\boldsymbol{\beta}}^{(\text{REWLS},\mathfrak{T})}_{nT} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n}\sum_{t=1}^{T^{(\mathfrak{T})}} \hat{w}_{it}\left(\hat{\boldsymbol{\beta}}^0_{nT}, \hat{\sigma}^0_{nT}; \hat{v}_{nT}\right) r^{2,(\mathfrak{T})}_{it}(\boldsymbol{\beta}). \tag{2.13}$$

This method is proved to preserve the breakdown-point properties of the initial robust estimator and achieve the asymptotic efficiency for Gaussian errors.

An alternative to the traditional one-step estimators is the reweighted least trimmed squares (RLTS) estimator (Čížek, 2010). Similarly to Gervini and Yohai (2002), weights (2.11) are constructed using the data-dependent cutoff point $\hat{v}_{nT}$. The resulting weights are however used within the LTS estimator rather than LS. Since LTS requires only the total number h_{nT} of observations to be included in the objective function, the number of observations with non-zero weights $\hat{w}_{it}(\cdot, \cdot; \hat{v}_{nT})$ has to be counted:

$$\hat{h}_{nT} = \sum_{i=1}^{n} \sum_{t=1}^{T^{(\mathfrak{T})}} I\left(\left| r_{it}^{(\mathfrak{T})}\left(\hat{\boldsymbol{\beta}}_{nT}^{0}\right) / \hat{\sigma}_{nT}^{0} \right| < \hat{v}_{nT}\right) = \sum_{i=1}^{n} \sum_{t=1}^{T^{(\mathfrak{T})}} \hat{w}_{it}\left(\hat{\boldsymbol{\beta}}_{nT}^{0}, \hat{\sigma}_{nT}^{0}; \hat{v}_{nT}\right). \tag{2.14}$$

The RLTS estimator is then simply defined as LTS using the data-dependent amount of trimming $\hat{h}_{nT}$ applied to the $\mathfrak{T}$-transformed panel data:

$$\hat{\boldsymbol{\beta}}_{nT}^{(\mathrm{RLTS},\mathfrak{T})} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{j=1}^{\hat{h}_{nT}} r_{(j)}^{2,(\mathfrak{T})}(\boldsymbol{\beta}). \tag{2.15}$$

Similarly to REWLS, RLTS preserves the breakdown-point properties of the initial robust estimator. Additionally, RLTS is asymptotically independent of the initial estimator and achieves asymptotic efficiency when errors are normally distributed.

Note that all discussed one-step estimators preserve the breakdown point of the initial LTS estimator. If LTS does not break down, that is, if the transformed data are in a general position and contain less than 50% outliers (see the proof of Theorem 1), the results of Gervini and Yohai (2002, Theorem 3.3) and Čížek (2010, Theorem 2) apply to the transformed samples and imply no breakdown for REWLS and RLTS, respectively. These different robust methods could differ though by the bias caused by outliers and in their finite-sample and asymptotic variances.

2.3.4 Asymptotic properties

The estimators introduced in the previous sections are applied to model (2.1) after the first-difference or pairwise-difference transformations, which lead to the serial correlation of the errors in (2.6) or (2.7), respectively. Almost all one-step or data-adaptive robust regression estimators are however asymptotically studied under the assumption of independent (and often identically distributed) errors, be it in the context of cross-sectional (Gervini and Yohai, 2002) or panel data (Lucas et al., 2007), or there are no asymptotic results available (Bramati and Croux, 2007). This limits also the extent to which we can characterize the asymptotic distribution of the proposed estimators. In particular, the asymptotic distribution under the first- and pairwise-differences can be easily derived only for the initial LTS estimator and its reweighted form RLTS (with the notable exception of the estimation based only on the first differences taken only at even time periods, which produces independent errors).

Now, the assumptions necessary to derive the asymptotic distribution of LTS and RLTS are presented. To this end, let $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT})^\top$, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})^\top$, $\tilde{\boldsymbol{X}}_i = (\tilde{\boldsymbol{x}}_{i1}, \ldots, \tilde{\boldsymbol{x}}_{iT^{(\mathfrak{T})}})^\top$, $\tilde{\boldsymbol{y}}_i = (\tilde{y}_{i1}, \ldots, \tilde{y}_{iT^{(\mathfrak{T})}})^\top$, and $\tilde{\varepsilon}_{it} = \tilde{y}_{it} - \tilde{\boldsymbol{x}}_{it}\boldsymbol{\beta}^0$ for all $i \in \mathbb{N}$ and $t = 1, \ldots, T^{(\mathfrak{T})}$, where $\boldsymbol{\beta}^0$ is the true parameter value in model (2.1). Further, let us recall that, in this context, $\lambda \in [1/2, 1]$ refers to the limits $\lim_{n\to\infty} h_{nT}/nT^{(\mathfrak{T})}$ or $\lim_{n\to\infty} \hat{h}_{nT}/nT^{(\mathfrak{T})}$, see (2.14), and that $T \geq 2$ is a fixed integer. The assumptions and the asymptotic distribution will be stated for symmetrically distributed errors for the sake of simplicity. The presented results can be derived analogously also under more general assumptions using Čížek (2010), where their detailed discussion can be found.

Assumption A

A1 Random vectors $\boldsymbol{y}_i$ and matrices $\boldsymbol{X}_i$ are independent and identically distributed for all $i \in \mathbb{N}$ and have finite second moments.

A2 Let $\{\varepsilon_{it}\}_{i\in\mathbb{N}}$ be a sequence of random variables with finite second moments and $\mathrm{E}(\varepsilon_{it}|\boldsymbol{X}_i) = 0$ for all $i \in \mathbb{N}$ and $t = 1, \ldots, T$. Further, the unconditional distribution function F of ε_{it} is assumed to be unimodal, absolutely continuous, and symmetrically distributed condionally on $\boldsymbol{X}_i$. Its density function has to be bounded and continuously differentiable.

A3 Let $Q(\lambda) = \mathrm{E}[\tilde{\boldsymbol{X}}_i^\top \operatorname{diag}(\{I[|F(\tilde{\varepsilon}_{it}) - F(-\tilde{\varepsilon}_{it} - 2C)| \leq \lambda]\}_{t=1}^{T^{(\mathfrak{T})}})\tilde{\boldsymbol{X}}_i]$ be a nonsingular matrix for any fixed $C \in \mathbb{R}$.

A4 Denoting G_β and g_β the unconditional cumulative distribution and density functions of $(\tilde{y}_{it} - \tilde{\boldsymbol{x}}_{it}^\top\boldsymbol{\beta})^2$, let $\sup_{\beta\in\mathbb{R}^p} \sup_{z>\alpha} g_\beta(z) < \infty$ for any $\alpha > 0$, and if $\lambda < 1$, that $\inf_{\beta\in\mathbb{R}^p} \inf_{z\in(-\delta,\delta)} g_\beta\left(G_\beta^{-1}(\lambda) + z\right) > 0$ for some $\delta > 0$.

Assumption A1 formulates standard conditions of the (uniform) central limit theorem: observed variables are independent across cross-sectional units and have finite second moments. Assumption A2 presents the assumptions on the error term ε_{it}, which is mean-independent of explanatory variables and continuously distributed. Note that, in the most general case, only the second moments of the trimmed errors $\boldsymbol{e}_i(q_\lambda)$ defined below have to be finite (see Čížek, 2011). Next, Assumption A3 formulates an analog of the standard full-rank condition and is actually equivalent to $\mathrm{E}(\tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{X}}_i^\top) > 0$ if $\tilde{\varepsilon}_{it}$ is independent of $\boldsymbol{X}_i$. Finally, Assumption A4 formalizes the fact that the distribution of squared residuals should be absolutely continuous with an everywhere finite density: its density function should not approach ∞ for any value of the parameter vector β. If $\tilde{\varepsilon}_{it}$ is independent of $\tilde{\boldsymbol{x}}_{it}$, Assumption A4 is usually implied by F being absolutely continuous with a density function f positive, bounded and differentiable (Čížek, 2006).

Under Assumption A, we can generalize the results of Čížek (2010) to the panel data model (2.1) and find the asymptotic distribution of LTS and RLTS. To formulate this result, the notation $q_\lambda = \sqrt{G^{-1}(\lambda)}$ is used, where $G \equiv G_\beta^0$ and G^{-1} represents the unconditional quantile function of $\tilde{\varepsilon}_{it}^2$. Additionally, one diagonal matrix and two vectors depending on q_λ are needed: $\boldsymbol{I}_i(q_\lambda) = \mathrm{diag}[\{I(\tilde{\varepsilon}_{it} \leq q_\lambda)\}_{t=1}^{T(\mathfrak{T})}]$, $\boldsymbol{e}_i(q_\lambda) = \boldsymbol{I}_i(q_\lambda)(\tilde{\varepsilon}_{i1}, \ldots, \tilde{\varepsilon}_{iT(\mathfrak{T})})^\top$, and $\boldsymbol{f}_i(q_\lambda) = (f_{i1}(q_\lambda), \ldots, f_{iT(\mathfrak{T})}(q_\lambda))^\top$, where f_{it} is the conditional distribution of $\tilde{\varepsilon}_{it}|\tilde{\boldsymbol{X}}_i$.

Theorem 2. *Let Assumption A hold. Next, let* $\Sigma(\lambda) = \mathrm{E}[\tilde{\boldsymbol{X}}_i^\top\{\boldsymbol{e}_i(q_\lambda)\boldsymbol{e}_i(q_\lambda)^\top\}\tilde{\boldsymbol{X}}_i]$, $Q(\lambda) = \mathrm{E}[\tilde{\boldsymbol{X}}_i^\top \boldsymbol{I}_i(q_\lambda)\tilde{\boldsymbol{X}}_i]$, $J(\lambda) = -\mathrm{E}[q_\lambda\tilde{\boldsymbol{X}}_i^\top \mathrm{diag}\{\boldsymbol{f}_i(-q_\lambda) + \boldsymbol{f}_i(q_\lambda)\}\tilde{\boldsymbol{X}}_i]$, *and* $Q(\lambda) + J(\lambda)$ *be a non-singular matrix. Then the (reweighted) LTS estimator* $\hat{\boldsymbol{\beta}}_{nT}^{(\mathrm{RLTS},\mathfrak{T})}$ *defined by trimming* $\hat{h}_{nT}$ *such that* $\lim_{n\to\infty}\hat{h}_{nT}/nT \to \lambda$ *for some* $\lambda \in \langle 1/2, 1\rangle$ *is a* $\sqrt{n}$*-consistent and asymptotically normal,* $\sqrt{n}(\hat{\boldsymbol{\beta}}_{nT}^{(\mathrm{RLTS},\mathfrak{T})} - \beta^0) \xrightarrow{\mathcal{L}} N(0, V(\lambda))$ *as* $n \to \infty$*, where the asymptotic covariance matrix equals* $V(\lambda) = \{Q(\lambda) + J(\lambda)\}^{-1}\Sigma(\lambda)\{Q(\lambda) + J(\lambda)\}^{-1}$.

The theorem covers not only the reweighted, but also the initial LTS estimator for $\hat{h}_{nT} = h_{nT} = \mathrm{const}$. Consequently, the initial and reweighted LTS estimators are asymptotically normal. The estimation of their covariance matrix $V(\lambda)$ is discussed in detail by Čížek (2011).

2.4 Simulation study

This section contains a simulation study of the finite-sample properties of some proposed and existing panel-data estimators. The following simulations are meant to investigate the behavior of estimators when the sample dimensions vary (Section 2.4.1), when errors come from various error distributions (Section 2.4.2), and when different kinds of outlying observations are present (Section 2.4.3). The reference estimator is the within-group estimator $\hat{\boldsymbol{\beta}}_{nT}^{(\mathrm{LS,\,mean})}$. Other estimators under consideration are the LS, LTS with the maximum amount of trimming (see Theorem 1), WGM and WMS of Bramati and Croux (2007), REWLS, and RLTS estimators subject to two data transformations $\mathfrak{T} \in \{\mathrm{med}, P\Delta\}$ (the first-difference transformation is omitted as it is inferior to the pairwise-difference transformation). Note that the LTS with the maximum amount of trimming serves as an initial estimator for all two-step estimators (i.e., for RLTS, REWLS, and WGM).

The data generating process is given by a static fixed-effect panel-data model

$$\begin{aligned} y_{it} &= \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + \alpha_i + \varepsilon_{it}, \\ \alpha_i &= \sum_{t=1}^{T} \boldsymbol{x}_{it}^\top \boldsymbol{\gamma}/\sqrt{T} + \eta_i, \end{aligned} \qquad i = 1, \cdots, n, \quad t = 1, \cdots, T, \tag{2.16}$$

where the ε_{it}'s are independent and identically distributed according to some distribution H. The parameters of interest are chosen $\boldsymbol{\beta} = (1, 0, -1)^\top$ (without loss of generality for all regression equivariant estimators). The unobservable individual effects α_i depend on $\eta_i \sim \mathrm{U}(0, 12)$ and on the covariates $\boldsymbol{x}_{it}$ through $\boldsymbol{\gamma} = (2, 2, 2)^\top$, where $\boldsymbol{\gamma}$ and the distribution of η_i are chosen so that the variance of the deterministic and random parts of α_i are equal. Observable covariates $\boldsymbol{x}_{it}$'s are generated according to

$$x_{itk} \sim \begin{cases} \chi_2^2 - 2 & \text{if } k = 1, \\ \mathrm{N}(0, 1) & \text{if } k \geq 2, \end{cases} \tag{2.17}$$

where x_{itk} denotes the kth component of $\boldsymbol{x}_{it}$, $k = 1, 2, 3$, χ_2^2 denotes the chi-squared distribution with 2 degrees of freedom, and $N(0, 1)$ represents the standard normal distribution. One asymmetrically distributed variable is included to avoid an 'ideal' completely symmetric design of the experiment.

Simulation experiments are conducted across different sample sizes nT, aiming at both short micro-panels and longer macro-panels, with n and T ranging from $(n, T) = (200, 3)$ to $(n, T) = (50, 24)$. The performance of each estimator is evaluated using $S = 1000$ simulated samples and is measured by the mean squared error (MSE): $MSE = 1/S \sum_{s=1}^{S} \|\hat{\boldsymbol{\beta}}_{nT}^s - \boldsymbol{\beta}\|^2$, where $\hat{\boldsymbol{\beta}}_{nT}^s$, $s = 1, \ldots, S$, are the estimates for S simulated samples. Additionally, we also report the power of the (simulated) tests of significance of individual coefficients $P_j = 1/S \sum_{s=1}^{S} I(|\hat{\beta}_{jnT}^s / \hat{s}_j^s| > z_{0.975})$, where $(\hat{s}_j^s)^2 = 1/S \sum_{s=1}^{S} (\hat{\beta}_{jnT}^s - \sum_{s=1}^{S} \hat{\beta}_{jnT}^s / S)^2$, $z_{0.975}$ is the 97.5% quantile of the standard

Table 2.2: The mean squared errors of all estimators for the linear panel-data model with normally distributed errors and various sample sizes.

n		50	100	200	50				
T			3		4	5	6	12	24
mean	LS	0.024	0.012	0.006	0.015	0.012	0.009	0.004	0.002
	WMS	0.043	0.019	0.009	0.022	0.017	0.012	0.005	0.002
	LS	0.045	0.032	0.025	0.019	0.022	0.012	0.006	0.003
	LTS	1.427	1.534	1.584	0.174	0.479	0.139	0.059	0.027
med	WGM	0.513	0.475	0.443	0.046	0.100	0.031	0.013	0.006
	REWLS	0.592	0.539	0.505	0.043	0.093	0.025	0.008	0.003
	RLTS	0.316	0.261	0.221	0.036	0.051	0.020	0.007	0.003
	LS	0.024	0.012	0.006	0.015	0.012	0.009	0.004	0.002
	LTS	0.138	0.075	0.044	0.079	0.058	0.042	0.013	0.005
$P\Delta$	WGM	0.032	0.015	0.008	0.019	0.014	0.011	0.005	0.002
	REWLS	0.031	0.013	0.006	0.017	0.013	0.010	0.004	0.002
	RLTS	0.028	0.013	0.006	0.016	0.012	0.009	0.004	0.002

normal distribution, and $j = 1, 2, 3$.

2.4.1 Sample sizes

The performance of the estimators is first evaluated for normal errors, $H \equiv N(0, 1)$, at different sample sizes: for T fixed and n increasing and for T increasing while n is fixed. The simulation results are summarized in Table 2.2 (note that the WMS estimator is included in the table section "median," as the program code of WMS supplied by the authors Bramati and Croux (2007) estimates the fixed effects by the median). The results for the median transformation confirm that the robust estimators based on this transformation are not consistent for a fixed number of time periods T, but are consistent if $T \to \infty$. Note that the median-transformation results are worse for the odd numbers of periods $T = 3, 5$ than for the even numbers of periods $T = 4, 6$ as the median equals to an average of two observations in the latter case. Such differences are becoming negligible for $T \geq 8$. On the contrary, the WMS estimator – despite estimating the fixed effects – exhibits small MSEs and performs similarly to or better than LS after the median transformation.

Next, LTS performs much worse than LS for both transformations, while all one-step estimators (WGM, REWLS, and RLTS) exhibit much smaller MSEs and can match the performance of LS if the sample size is

Table 2.3: The mean squared errors of all estimators for the linear panel-data model with errors from the standard normal, double exponential, and Student distributions.

Errors distr.		DExp(1)			N(0,1)			t_3		
n		200	75	30	200	75	30	200	75	30
T		3	8	20	3	8	20	3	8	20
mean	LS	0.011	0.008	0.008	0.006	0.004	0.004	0.016	0.013	0.012
	WMS	0.014	0.009	0.009	0.009	0.005	0.005	0.019	0.013	0.012
med	LS	0.032	0.011	0.008	0.026	0.007	0.005	0.038	0.015	0.013
med	LTS	1.643	0.074	0.034	1.587	0.085	0.050	1.656	0.088	0.049
med	WGM	0.511	0.023	0.012	0.443	0.018	0.010	0.520	0.025	0.013
med	REWLS	0.582	0.018	0.009	0.505	0.011	0.006	0.593	0.019	0.009
med	RLTS	0.242	0.016	0.009	0.221	0.010	0.006	0.250	0.016	0.009
$P\Delta$	LS	0.011	0.008	0.008	0.006	0.004	0.004	0.016	0.013	0.012
$P\Delta$	LTS	0.034	0.014	0.008	0.043	0.018	0.011	0.044	0.020	0.011
$P\Delta$	WGM	0.010	0.007	0.006	0.008	0.005	0.004	0.012	0.008	0.007
$P\Delta$	REWLS	0.010	0.007	0.006	0.006	0.004	0.004	0.012	0.008	0.007
$P\Delta$	RLTS	0.010	0.008	0.007	0.006	0.004	0.004	0.012	0.008	0.007

sufficiently large. Additionally, the REWLS and RLTS estimators, which are asymptotically equivalent to LS, produce the same MSEs as the within-group estimator in the samples of 600 or more observations.

2.4.2 Different error distributions

In this subsection, three different distributions H of the error term ε_{it} in (2.16) are considered: the standard normal $N(0,1)$, the double exponential distribution $DExp(1)$ with rate 1, and the Student distribution t_3 with 3 degrees of freedom, see Table 2.3. The within-group LS estimator is no longer optimal (although it performs better than WMS) and is slightly outperformed by one-step robust estimators based on the pairwise differences in the case of the double-exponential errors and more substantially in the case of the Student errors (the differences among WGM, REWLS, and RLTS are practically negligible).

2.4.3 Outliers

The robust properties are now evaluated by including outliers in the data. Let m be the number of outliers and let I_m be the index set of contaminated observations. Contaminated values of the dependent variable $\breve{y}^r_{it} \sim U(-10,30)$ or $\breve{y}^c_{it} \sim U(29,30)$ and independent variables $\breve{x}_{itk} \sim N(6,2)$, $(i,t) \in I_m$, $k = 1,2,3$, result in the following contamination schemes defined by the actual values of $(\boldsymbol{x}_{it}, y_{it})$ for $(i,t) \in I_m$. If $y_{it} = \breve{y}^r_{it}$ or $y_{it} = \boldsymbol{x}^\top_{it}\boldsymbol{\beta} + \alpha_i + \breve{y}^c_{it}$ for $(i,t) \in I_m$, we talk about the non-clustered and clustered outliers, respectively. On the other hand, if $\boldsymbol{x}_{it}$ is left unmodified or $\boldsymbol{x}_{it} = \breve{\boldsymbol{x}}_{it}$, the contamination schemes are said to contain vertical outliers (VO) or leverage points (LP), respectively. All non-contaminated data $(i,t) \notin I_m$ follow model (2.16)–(2.17) with $H \equiv N(0,1)$. The total sample size is 210 and is attained by setting $n = 105$ and $T = 2$ or $n = 70$ and $T = 3$ and the number of outliers is set to $m = 10$ (5% contamination) and $m = 42$ (20% contamination). Note that outliers are distributed randomly across cross-sectional units. We therefore report $T = 2$ and $T = 3$ to demonstrate the effect of outliers concentrated within individual units: the fraction of outliers per one cross-sectional unit is always at least one half ($T = 2$) or not ($T = 3$). This is largely irrelevant for the proposed estimators, but matters for the WGM and WMS estimators.

The results summarized in Tables 2.4–2.6. The first one documents that, even if only 5% observations are contaminated, LS can get extremely biased (actually more than the inconsistent estimators based on the median transformation). Although the median-based WGM is biased for $T = 3$ to the same extent as in the previous experiments without outliers, it delivers reliable estimates for $T = 2$. On the other hand, the robust WMS estimator provides robust results at 5% contamination, but can be severely affected by 20% contamination if $T = 2$: the large MSE of WMS likely stems from the fact that the estimation of fixed effects for any contaminated individual is substantially biased if $T = 2$ (or in general, if outliers tend to be concentrated within particular individuals as discussed by Bramati and Croux (2007)). On the contrary, the proposed robust estimators are not substantially affected by any type of contamination. Similarly to experiments discussed in previous sections, there are no substantial differences among the one-step robust estimators.

The next Tables 2.5–2.6 display the biases and the simulated powers and sizes of the significance tests for the coefficients $\beta_1 = 1$ and $\beta_2 = 0$, respectively, in the presence of 20% outliers. In the case of LS, one observes that the vertical outliers lead to the loss of power due to large variances, whereas the leverage points cause substantial biases. The biases and loss of power are also visible in the case of WMS for $T = 2$ with leverage points. The median-based WGM method also exhibits substantial biases, but for $T = 3$. Finally, the proposed estimators based on the pairwise difference transformation do not exhibit substantial biases (the "worst" cases

Table 2.4: The mean squared errors of all estimators for the linear panel-data models with $T = 2$ and $T = 3$ in the presence of 5% or 20% scattered and clustered outliers coming from $U(-10, 30)$ and $U(29, 30)$ distributions. Contamination denoted 'VO' and 'LP' refers to outlying observations being vertical outliers and leverage points, respectively.

(n, T)			Non-clustered outliers				Clustered outliers			
			VO		LP		VO		LP	
			5%	20%	5%	20%	5%	20%	5%	20%
$(105, 2)$	mean	LS	0.247	0.961	3.733	5.663	1.032	3.338	6.262	8.096
		WMS	0.053	0.185	0.067	2.684	0.056	2.247	0.060	8.042
		LS	0.247	0.961	3.733	5.663	1.032	3.338	6.262	8.096
		LTS	0.158	0.123	0.160	0.140	0.162	0.094	0.169	0.087
	med	WGM	0.037	0.065	0.040	0.093	0.033	0.037	0.035	0.038
		REWLS	0.035	0.068	0.044	0.116	0.030	0.036	0.032	0.039
		RLTS	0.033	0.073	0.040	0.186	0.029	0.036	0.030	0.039
		LS	0.251	0.968	3.738	5.668	1.046	3.379	6.262	8.094
		LTS	0.151	0.122	0.163	0.137	0.161	0.090	0.168	0.087
	$P\Delta$	WGM	0.037	0.066	0.041	0.093	0.033	0.037	0.035	0.038
		REWLS	0.037	0.069	0.046	0.115	0.032	0.038	0.034	0.041
		RLTS	0.034	0.074	0.042	0.171	0.031	0.037	0.031	0.040
$(70, 3)$	mean	LS	0.171	0.654	2.585	5.312	0.654	2.331	6.040	8.017
		WMS	0.061	0.107	0.068	0.118	0.067	0.093	0.070	0.088
		LS	0.182	0.608	3.112	5.842	0.554	1.903	6.314	8.243
		LTS	1.521	1.642	1.466	1.401	1.494	1.559	1.546	1.509
	med	WGM	0.491	0.562	0.482	0.474	0.466	0.431	0.452	0.444
		REWLS	0.572	0.653	0.557	0.530	0.529	0.479	0.553	0.514
		RLTS	0.262	0.252	0.288	0.309	0.261	0.223	0.216	0.327
		LS	0.171	0.658	2.584	5.316	0.658	2.346	6.041	8.017
		LTS	0.097	0.078	0.102	0.086	0.098	0.050	0.098	0.045
	$P\Delta$	WGM	0.023	0.042	0.026	0.054	0.022	0.022	0.022	0.021
		REWLS	0.021	0.045	0.032	0.077	0.020	0.023	0.019	0.021
		RLTS	0.021	0.048	0.030	0.136	0.020	0.022	0.019	0.021

Table 2.5: The biases (upper rows) and powers (lower rows) of all estimators for the linear panel-data models with $T = 2$ in the presence of 20% scattered and clustered outliers coming from $U(-10, 30)$ and $U(29, 30)$ distributions. Contamination denoted 'VO' and 'LP' refers to outlying observations being vertical outliers and leverage points, respectively.

$(n,T) = (105, 2)$		Non-clustered outliers				Clustered outliers			
		VO		LP		VO		LP	
		β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
mean	LS	0.202	0.028	-0.172	-1.411	0.006	0.018	0.659	1.931
		0.730	0.052	0.942	0.760	0.364	0.047	0.412	1.000
	WMS	0.019	0.010	-0.228	-0.992	0.009	-0.004	0.674	1.919
		0.999	0.053	1.000	0.416	0.470	0.063	0.286	0.998
	LS	0.202	0.028	-0.172	-1.411	0.006	0.018	0.659	1.931
		0.730	0.052	0.942	0.760	0.364	0.047	0.412	1.000
	LTS	0.002	0.007	-0.015	-0.049	-0.005	-0.008	0.002	-0.000
		1.000	0.052	1.000	0.061	1.000	0.056	0.999	0.046
med	WGM	0.000	0.007	-0.035	-0.111	-0.002	-0.005	0.002	-0.004
		1.000	0.053	1.000	0.096	1.000	0.054	0.999	0.031
	REWLS	-0.000	0.006	-0.033	-0.117	-0.002	-0.005	0.002	-0.004
		1.000	0.057	1.000	0.090	1.000	0.050	0.999	0.033
	RLTS	-0.000	0.007	-0.051	-0.187	-0.002	-0.005	0.002	-0.004
		1.000	0.057	1.000	0.126	1.000	0.050	0.999	0.034
	LS	0.201	0.030	-0.172	-1.412	0.008	0.013	0.660	1.931
		0.721	0.055	0.941	0.752	0.361	0.048	0.413	1.000
	LTS	-0.000	0.007	-0.016	-0.054	-0.003	-0.007	0.001	-0.001
		1.000	0.051	1.000	0.060	1.000	0.053	0.999	0.045
$P\Delta$	WGM	0.000	0.007	-0.036	-0.113	-0.002	-0.005	0.002	-0.004
		1.000	0.052	1.000	0.102	1.000	0.054	0.999	0.030
	REWLS	0.000	0.007	-0.032	-0.113	-0.003	-0.004	0.003	-0.003
		1.000	0.061	1.000	0.087	1.000	0.051	0.999	0.031
	RLTS	0.000	0.008	-0.047	-0.175	-0.002	-0.004	0.003	-0.003
		1.000	0.051	1.000	0.119	1.000	0.050	0.999	0.034

are REWLS and RLTS in the case with non-clustered leverage points) and the test sizes and powers P_j are close to the ideal values 0.05 and 1.00, respectively, in almost all cases (the exception is RLTS in the case of non-clustered leverage points).

2.5 An empirical application

In this section we illustrate the proposed methods by means of an empirical application. We consider the wage equation example as studied in Cornwell and Rupert (1988). The data set is drawn from the Panel Study of Income Dynamics (PSID) and consists of 595 individuals observed in years 1976–1982 (i.e., $T = 7$). In particular, the logarithm of wage is regressed on

Table 2.6: The biases (upper rows) and powers (lower rows) of all estimators for the linear panel-data models with $T = 3$ in the presence of 20% scattered and clustered outliers coming from $U(-10, 30)$ and $U(29, 30)$ distributions. Contamination denoted 'VO' and 'LP' refers to outlying observations being vertical outliers and leverage points, respectively.

$(n, T) = (70, 3)$		Non-clustered outliers				Clustered outliers			
		VO		LP		VO		LP	
		β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
mean	LS	0.200	-0.010	-0.201	-1.359	-0.001	0.002	0.653	1.929
		0.881	0.058	0.973	0.792	0.471	0.044	0.538	1.000
	WMS	0.003	-0.004	-0.013	-0.054	0.009	-0.003	0.011	0.020
		1.000	0.073	1.000	0.083	1.000	0.070	1.000	0.070
	LS	0.260	-0.013	-0.058	-1.359	0.101	-0.012	0.811	1.965
		0.899	0.054	0.951	0.798	0.484	0.051	0.173	1.000
	LTS	0.891	-0.002	0.712	-0.012	0.848	0.001	0.929	0.046
		0.067	0.055	0.159	0.070	0.084	0.052	0.037	0.124
med	WGM	0.496	-0.007	0.316	-0.072	0.409	0.001	0.581	0.165
		0.996	0.056	0.997	0.148	1.000	0.044	0.999	0.479
	REWLS	0.563	-0.004	0.363	-0.057	0.459	-0.000	0.623	0.154
		0.857	0.050	0.961	0.089	0.973	0.052	0.971	0.377
	RLTS	0.216	-0.007	0.125	-0.074	0.191	0.000	0.453	0.175
		0.986	0.052	0.999	0.090	0.998	0.049	0.760	0.372
	LS	0.201	-0.009	-0.201	-1.361	-0.000	0.001	0.653	1.928
		0.883	0.062	0.973	0.791	0.468	0.043	0.534	1.000
	LTS	-0.006	0.005	-0.007	-0.014	0.003	-0.004	-0.000	-0.002
		1.000	0.047	1.000	0.053	1.000	0.053	1.000	0.050
$P\Delta$	WGM	0.000	-0.004	-0.017	-0.060	0.001	-0.001	0.002	-0.003
		1.000	0.057	1.000	0.068	1.000	0.049	1.000	0.047
	REWLS	-0.001	-0.004	-0.015	-0.065	0.002	-0.001	0.001	-0.003
		1.000	0.050	1.000	0.064	1.000	0.048	1.000	0.051
	RLTS	0.001	-0.006	-0.032	-0.129	0.002	-0.001	0.002	-0.003
		1.000	0.052	1.000	0.106	1.000	0.046	1.000	0.051

Table 2.7: Fixed-effects estimates and their standard errors (in brackets) for the wage data from the Panel Study of Income Dynamics in the period 1976–1982. Dependent variable is the logarithm of wage. Note: the results for the WMS estimator are taken from Baltagi and Bresson (2011).

Transformation	mean	med		$P\Delta$		
Estimator:	LS	WMS	WGM	LTS	REWLS	RLTS
EXPSQ	-0.0004	-0.0005	-0.0003	-0.0002	-0.0004	-0.0004
	(0.0001)	—	—	(0.0008)	—	(0.0000)
EXP	0.1132	0.1105	0.1015	0.0982	0.1058	0.1084
	(0.0025)	—	—	(0.0409)	—	(0.0019)
WKS	0.0008	0.0020	-0.0006	-0.0003	0.0009	0.0013
	(0.0006)	—	—	(0.0046)	—	(0.0004)
OCC	-0.0215	-0.0224	-0.0037	-0.0023	-0.0172	-0.0237
	(0.0138)	—	—	(0.2420)	—	(0.0107)
IND	0.0192	-0.0177	-0.0080	0.0161	0.0043	0.0029
	(0.0154)	—	—	(0.3708)	—	(0.0119)
SOUTH	-0.0019	-0.0975	-0.0617	-0.0654	-0.0398	-0.0139
	(0.0343)	—	—	(0.1488)	—	(0.0316)
SMSA	-0.0425	-0.0209	-0.0137	-0.0328	-0.0201	-0.0162
	(0.0194)	—	—	(0.1001)	—	(0.0164)
MS	-0.0297	0.0345	-0.0185	-0.0006	-0.0155	-0.0202
	(0.0190)	—	—	(0.4369)	—	(0.0142)
UNION	0.0328	0.0363	0.0101	0.0074	0.0109	0.0073
	(0.0149)	—	—	(0.1043)	—	(0.0123)

the available time-varying variables: work experience (EXP), weeks worked (WKS), occupation (OCC $=1$ if the individual has blue-collar occupation), industry (IND $=1$ if the individual works in a manufacturing industry), residence (SOUTH $=1$ and SMSA $=1$ if the individual resides in the South or in a standard metropolitan area, respectively), marital status (MS $=1$ if the individual is married), and the union coverage (UNION $=1$ if an individual's wage is set by a union contract); see Cornwell and Rupert (1988) for a more detailed description of the data set. Note that, given the presence of many explanatory variables (including categorical ones), this empirical application seems to be sufficiently complex to analyze the applicability of the proposed estimators to real applications, in particular of the LTS and RLTS estimators. (The LTS estimation routine has been adjusted so that the subsampling algorithm does not break down in samples with many discrete variables.)

Results are reported in Table 2.7, where the standard errors represent the asymptotic standard deviations of estimates computed using Theorem 2 under the assumption of homoscedasticity. The estimated percentages of adaptive trimming for REWLS and RLTS are roughly equal to 10%, that is, $h_{nT}/nT^{(\mathfrak{T})} \approx 90\%$. For those estimators for which we compute standard errors, RLTS yields more accurate estimates than the within-group LS estimator in all cases (while LTS exhibits large standard errors). The signs and magnitudes of estimates are similar across LS, WMS, REWLS, and RLTS for several variables such as the experience and its square. On the other hand, the conclusions differ significantly for other variables. For example, the number of weeks worked WKS does not seem to be significant given the LS estimates, but it is significantly positive given the RLTS output; the local dummy SMSA is significant in LS, but is insignificant in RLTS as the robustly estimated coefficient is almost three times smaller in absolute value than the LS coefficient. Finally, it is interesting to note that the marriage variable MS seems to have an expected negative effect on the wages judging by the LS and all one-step robust estimates, but WMS estimates a positive effect of this variable. We can thus conclude that the proposed estimators, in particular RLTS here, provide both more robust and more precise estimates.

2.6 Concluding remarks

The present study examines the parameter estimation in fixed-effects panel data models with a fixed number of time periods from the point of view of robust statistical procedures. To achieve consistent estimators, we propose the pairwise-difference data transformation and then apply two robust estimators: LTS followed by various data-adaptive reweighted LS and LTS methods. For a given data transformation, all methods achieve the same breakdown point and have similar finite-sample performance; the asymptotic distribution could be however provided only in the case of LTS and RLTS. The benefits of the pairwise-difference transformation are, for example, the breakdown point 1/4 or more irrespective of the number of time periods T and good finite-sample performance of robust estimators based on this transformation. This could motivate its further study in the context of panel data models as it can be readily applied to many extensions of the standard fixed-effect panel data models, such as those discussed in Coakley et al. (2006).

Acknowledgements

This research was partially supported by grant GA ČR 402/09/0557. We would also like to thank to the anonymous referees for many useful comments, which helped us to substantially improve this work.

2.A Proofs

Proof of Lemma 1: Since the LTS estimator is regression, affine, and scale equivariant (Rousseeuw and Leroy, 1987, Lemma 3 in Chapter 3), we only have to verify that the data-transformations – the first- and pairwise-differencing – do not affect the regression, affine, and scale transformations. For any $s \in N$, this directly follows from $\Delta^s(cy_{it}) = c\Delta^s y_{it}$, $\Delta^s(y_{it} + \boldsymbol{x}_{it}^\top \boldsymbol{v}) = \Delta^s y_{it} + (\Delta^s \boldsymbol{x}_{it})^\top \boldsymbol{v}$, and $\Delta^s(\boldsymbol{x}_{it}^\top \boldsymbol{A}) = (\Delta^s \boldsymbol{x}_{it})^\top \boldsymbol{A}$. □

Proof of Theorem 1: Before applying the LTS estimator, data are subject to the differencing transformations (2.6) or (2.7), which generate $nT^{(\mathfrak{T})} = n(T-1)$ or $nT^{(\mathfrak{T})} = nT(T-1)/2$ transformed observations, respectively. With these transformations, the worst case scenario occurs when aberrant observations are located so that each single outlier contaminates always $\min\{2, T-1\}$ (first-differencing) or $T-1$ (pairwise-differencing) differentiated observations. Hence given m outliers in the original sample, the number of outliers after the first- and pairwise-differencing will be at most $\min\{2, T-1\}m$ and $(T-1)m$, respectively.

At the same time, the breakdown point of LTS with the trimming constant h_{nT} equals $(nT^{(\mathfrak{T})} - h_{nT})/[nT^{(\mathfrak{T})}]$ if $h_{nT} \geq (nT^{(\mathfrak{T})} + p + 1)/2$ (Vandev and Neykov, 1998). LTS thus breaks down only if the number of outliers exceeds $nT^{(\mathfrak{T})} - h_{nT}$. In the case of the first differences, this means that LTS breaks down if $\min\{2, T-1\}m > nT^{(\mathfrak{T})} - h_{nT}$, implying that the breakdown point of the proposed panel-data LTS estimator is greater or equal to $(nT^{(\mathfrak{T})} - h_{nT})/[\min\{2, T-1\}nT] = \{(nT^{(\mathfrak{T})} - h_{nT})/[2nT^{(\mathfrak{T})}]\} \cdot \{(2(T-1))/(\min\{2, T-1\}T)\}$. The limit claim of the theorem follows from $\lim_{n\to\infty}(nT^{(\mathfrak{T})} - h_{nT})/(2nT^{(\mathfrak{T})}) = (1-\lambda)/2$.

In the case of the pairwise differences, each outlier can contaminate at most $T-1$ pairwise differences. If this is true for every outliers, LTS breaks down if $(T-1)m > nT^{(\mathfrak{T})} - h_{nT}$, implying that the breakdown point equals $(nT^{(\mathfrak{T})} - h_n^{\mathfrak{T}})/[nT(T-1)] = (nT^{(\mathfrak{T})} - h_{nT})/(2nT^{(\mathfrak{T})})$ (in general, the breakdown point will be larger or equal to this bound).

The number of pairwise differences contaminated by one outlier decreases as the number of outliers per one cross-sectional unit increases. Thus, m outliers contaminate the largest number amount of pairwise differences if each cross-sectional unit contains the same amount m_1 of outliers (or possibly $m_1 - 1$ if m is not a multiple of n). The number of contaminated differences per unit is then at most $(T-1)+\ldots+(T-m_1) = [T(T-1)-(T-m_1)(T-m_1-1)]/2$, and if its n-multiple is larger than $nT^{(\Sigma)} - h_{nT} = nT(T-1)/2 - h_{nT}$, the LTS estimator breaks down. For a given h_{nT}, LTS thus breaks down if $h_{nT} > n(T-m_1)(T-m_1-1)/2$, where $m_1 = \lceil m/n \rceil$. Solving for the largest m and m_1 satisfying no-breakdown condition $2h_{nT}/n \le (T-m_1)(T-m_1)-(T-m_1)$ leads to $T - m_1 \ge 1/2 + \sqrt{2h_{nT}/n + 1/4}$, $m_1 \le T - 1/2 - \sqrt{2h_{nT}/n + 1/4}$, and $m \le n\lfloor T - 1/2 - \sqrt{2h_{nT}/n + 1/4} \rfloor$. Hence, LTS does not break down if $m/nT \le \lfloor T - 1/2 - \sqrt{2h_{nT}/n + 1/4} \rfloor / T$, where the right-hand side converges to $\lfloor T - 1/2 - \sqrt{2\lambda T^{(\Sigma)} + 1/4} \rfloor / T$ for $n \to \infty$. □

Proof of Theorem 2: Assumption A allows us to apply the results of Čížek (2010) to RLTS. Its objective function can be almost surely written as, see Čížek (2006, Lemma 1),

$$S_n(\beta) = \sum_{t=1}^{T^{(\Sigma)}} \sum_{i=1}^{n} r_{it}^{(\Sigma),2}(\beta) \cdot I\{r_{it}^{(\Sigma),2}(\beta) \le r_{[\hat{h}_n]}^{(\Sigma),2}(\beta)\}.$$

As the sequence $r_i^2(\beta)$ is $T^{(\Sigma)}$-dependent, the proof of consistency in Čížek (2010, Theorem 3) applies also to RLTS estimate $\hat{\beta}_n^{(RLTS,\Sigma)}$.

To derive the asymptotic distribution, note that the results for the cross-sectional LTS estimator of Čížek (2010) can be applied for any fixed $t = 1, \ldots, T^{(\Sigma)}$ to

$$S_{nt}(\beta) = \sum_{i=1}^{n} r_{it}^{(\Sigma),2}(\beta) \cdot I\{r_{it}^{(\Sigma),2}(\beta) \le r_{[\hat{h}_n]}^{(\Sigma),2}(\beta)\}$$

For example, the first-order conditions for minimizing $S_{nt}(\beta)$ following from Čížek (2010, equation (16) and Theorem 4) are

$$0 = S_{nt}^{'}(\beta^0) - 2\{Q_t(\lambda) + J_t(\lambda)\}\sqrt{n}(\hat{\beta}_n^{(RLTS,\Sigma)} - \beta^0) + o_p(1)$$

as $n \to \infty$, where $S_{nt}^{'}(\beta)$ denotes the derivative wrt. the parameter vector β, $Q_t(\lambda) = \mathrm{E}[\tilde{\boldsymbol{x}}_{it}\tilde{\boldsymbol{x}}_{it}^\top I(r_{it}^{(\Sigma),2}(\beta^0) \le q_\lambda^2)]$, and $J_t(\lambda) = \mathrm{E}[-\tilde{\boldsymbol{x}}_{it}\tilde{\boldsymbol{x}}_{it}^\top q_\lambda(f_{it}(q_\lambda) +$

$f_{it}(-q_\lambda))]$. Consequently,

$$0 = S_n^{'}(\beta^0) - 2\sum_{t=1}^{T^{(\mathfrak{T})}}\{Q_t(\lambda) + J_t(\lambda)\}\sqrt{n}(\hat{\beta}_n^{(RLTS,\mathfrak{T})} - \beta^0) + o_p(1).$$

Since $\sum_{t=1}^{T^{(\mathfrak{T})}}\{Q_t(\lambda) + J_t(\lambda)\} = Q(\lambda) + J(\lambda)$, we only have to derive the the asymptotic distribution of $S_n^{'}(\beta^0)/2$ as rewriting yields

$$\sqrt{n}(\hat{\beta}_n^{(RLTS,\mathfrak{T})} - \beta^0) = \{Q(\lambda) + J(\lambda)\}^{-1}S_n^{'}(\beta^0)/2 + o_p(1). \tag{2.18}$$

Because $S_n^{'}(\beta^0) = (\boldsymbol{I}_p, \ldots, \boldsymbol{I}_p)(S_{n1}^{'}(\beta^0)^\top, \ldots, S_{n\mathfrak{T}}^{'}(\beta^0)^\top)^\top$, where the last vector is independent and identically distributed, we can apply the central limit theorem to $S_n^{'}(\beta^0)/2$ as in the proof of Čížek (2010, Theorem 3), from equation (18) on, to find out that $S_n^{'}(\beta^0)/2$ is asymptotically normally distributed with variance matrix

$$(\boldsymbol{I}_p, \ldots, \boldsymbol{I}_p)\,\mathrm{var}\left[\{\mathrm{diag}(\boldsymbol{e}_i(q_\lambda)) \otimes \boldsymbol{I}_p\}\mathrm{vec}(\tilde{\boldsymbol{X}}_i^\top)\right](\boldsymbol{I}_p, \ldots, \boldsymbol{I}_p)^\top,$$

which is just another (more complex) form of $\Sigma(\lambda)$. The theorem follows from (2.18). $\square$

Chapter 3

Robust estimation of dynamic fixed-effects panel data models[1]

3.1 Introduction

In this paper, the robust estimation of dynamic panel data models with fixed effects is considered, which have proven to be very attractive models in empirical applications. An important advantage of these models is that they allow to disentangle the persistent component due to the (time-invariant) unobserved heterogeneity from the one based on the dynamic behavior. The related literature is fairly extensive and dates back to more than sixty years ago — for an overview, see among others Harris et al. (2008). Unfortunately, almost all literature focuses on the models assuming that data are free of outlying or aberrant observations. This is often not the case in reality, not even in relatively reliable macroeconomic data as documented in Zaman et al. (2001). This issue is even more important in the case of panel data, where erroneous observations can be masked by the complex structure of the data.

Despite its relevance, the study of robust techniques for panel data seems to be rather limited. Few contributions are available for static models and even fewer for the dynamic setting. Lucas et al. (2007) concentrates on constructing the generalized method of moment estimator with a bounded influence function. Galvao (2011) proposes to estimate the model using quantile regression techniques. Both these procedures focus on methods that are only locally robust. On the contrary, Dhaene and Zhu (2009) propose a

[1] This chapter is based on Aquaro M., and P. Čížek (2012), Robust estimation of dynamic fixed-effects panel data models, *Working Paper*.

median-based robust estimator that is globally robust and has sufficiently good behavior in non-contaminated data. This estimator is based on the median ratio of two consecutive pairs of the first-differenced data.

The aim of this paper is to extend the median-based estimator of Dhaene and Zhu (2009) by means of the multiple pairwise difference transformation to obtain a robust estimation with as good finite-sample performance as the commonly used non-robust estimators such as the one by Blundell and Bond (1998). The pairwise difference transformation has been previously used in the robust statistics, for example, by Stromberg et al. (2000) and Aquaro and Čížek (2013), who first create all possible pairwise differences of the available data and then apply a robust estimation technique on the set of all differences. Contrary to those existing applications, we apply the pairwise differencing on dependent data (time series), which prevents a straightforward application of the concept. Therefore, we first generalize the results of Dhaene and Zhu (2009) for a generic sth difference transformation, $s \in N$. Next, we combine multiple pairwise differences by means of the generalized method of moments (GMM) to estimate the parameter of interest more precisely. Finally, the robustness properties of the transformation for various sth differences are studied and used to choose the ones that do not lessen the robustness of the original estimator.

The paper is organized as follows. In Section 3.2, the new estimator in introduced and its asymptotic distribution is derived. Its robust properties are studied in Section 3.3. We present results of the Monte Carlo simulations in Section 3.4. The proofs are in the Appendix.

3.2 Estimator

Consider the dynamic panel data model

$$y_{it} = \alpha y_{it-1} + \eta_i + \varepsilon_{it} \qquad (t = 1, \ldots, T; \quad i = 1, \ldots, n), \tag{3.1}$$

where y_{it} is the response variable, η_i is the unobservable fixed effect, and ε_{it} represents the idiosyncratic error. To guarantee the stationarity of the data following the model, $|\alpha| < 1$ is assumed. The time dimension T can be fixed and possibly rather small (i.e., it does not have to increase with the sample size), and consequently, fixed or stochastic effects η_i are nuisance parameters, which cannot be consistently estimated. We concentrate on the estimation of this simple dynamic model as the main difficulty lies in the estimation of the autoregressive parameter α and the extension of the discussed estimators to a model including exogenous covariates is straightforward (see Dhaene and Zhu, 2009, Section 4.1).

We will consider model (3.1) under the following assumptions:

A.1 Errors ε_{it} are assumed to be independent across $i = 1, \ldots, n$ and $t = 1, \ldots, T$ and to possess finite second moments. Errors $\{\varepsilon_{it}\}_{t=1}^{T}$ are also independent of the fixed effects η_i.

A.2 The sequences $\{y_{it}\}_{t=1}^{T}$ are time stationary for all $i = 1, \ldots, n$. In particular, the first and second moments of y_{it} conditional to η_i do not depend of time.

A.3 Let $\varepsilon_{it} \sim \mathrm{N}(0, \sigma_i^2)$ for all $i = 1, \ldots, n$ and $t = 1, \ldots, T$.

First, note that no assumptions are made about the unobservable fixed effects η_i except for A.1. The errors ε_{it} are also not required to follow the same distribution across cross-sectional units i. Although we derive the results under the normality of the errors, see Assumption A.3, the estimator is consistent as long as the joint distribution of errors $\{\varepsilon_{it}\}_{t=1}^{T}$ is elliptically contoured (see Dhaene and Zhu, 2009, Section 4.2). Next, the stationarity Assumption A.2 is used not only by the discussed robust estimators, but also by frequently applied GMM estimators such as Blundell and Bond (1998). Finally, despite the fact that asymptotic properties are derived under heteroskedasticity, the assumption of homoskedasticity of the error $\sigma_i^2 = \sigma^2$ for $i = 1, \cdots, n$ will be needed to derive the robustness properties in Section 3.3.

The generalization of the estimator by Dhaene and Zhu (2009) will now be derived. Let Δ^s denote the sth difference operator, that is, $\Delta^s \upsilon_t := \upsilon_t - \upsilon_{t-s}$ (cf. Abrevaya, 2000; Aquaro and Čížek, 2013). Given model (3.1), it holds under stationarity for $s, q, p \in N$ that

$$\mathrm{E}(\Delta^s y_{it} | \Delta^p y_{it-q}) = r_{\boldsymbol{j}} \Delta^p y_{it-q}, \tag{3.2}$$

where the triplet $\boldsymbol{j} = (s, q, p)'$ and $r_{\boldsymbol{j}}$ are independent of i and t, $\max\{s, p + q\} < T$, and

$$r_{\boldsymbol{j}} := \frac{\mathrm{cov}(\Delta^s y_{it}, \Delta^p y_{it-q})}{\mathrm{var}(\Delta^p y_{it-q})}; \tag{3.3}$$

see for instance Bain and Engelhardt (1992, Theorem 5.4.6). For the sake of simplicity, we will assume that $s < p + q$.

Next, equation (3.2) implies that the variables $\Delta^s y_{it} - r_{\boldsymbol{j}}\Delta^p y_{it-q}$ and $\Delta^p y_{it-q}$ are uncorrelated, and by Assumption A.3, that they are independent and symmetrically distributed around zero. It follows that

$$\mathrm{E}[\mathrm{sgn}(\Delta^s y_{it} - r_{\boldsymbol{j}}\Delta^p y_{it-q})\,\mathrm{sgn}(\Delta^p y_{it-q})] = 0, \tag{3.4}$$

which can be rewritten more conveniently as

$$\mathrm{E}\left[\mathrm{sgn}\left(\frac{\Delta^s y_{it}}{\Delta^p y_{it-q}} - r_{\boldsymbol{j}}\right)\right] = 0. \tag{3.5}$$

This facilitates the estimation of $r_{\boldsymbol{j}}$ by the sample analog of this condition:

$$\hat{r}_{n\boldsymbol{j}} = \mathrm{med}\left\{\frac{\Delta^s y_{it}}{\Delta^p y_{it-q}};\ t = p+q+1,\dots,T;\ i = 1,\dots,n\right\}. \tag{3.6}$$

To relate this estimator to the autoregressive coefficient α, it is possible to derive under Assumption A.1–A.2 that the correlation coefficient $r_{\boldsymbol{j}}$ in (3.3) is equal to

$$r_{\boldsymbol{j}} = \frac{\alpha^q - \alpha^{q+p} - \alpha^{|s-q|} + \alpha^{|s-p-q|}}{2(1-\alpha^p)} \tag{3.7}$$

(see its derivation in the Appendix 3.A). This last equality can be rewritten more conveniently as

$$g_{\boldsymbol{j}}(\alpha) = 2(1-\alpha^p)r_{\boldsymbol{j}} - \alpha^q + \alpha^{q+p} + \alpha^{|s-q|} - \alpha^{|s-p-q|} = 0. \tag{3.8}$$

Dhaene and Zhu (2009) propose to estimate α by transforming (3.1) using the first difference, that is, by setting s, q, and p in (3.5) all equal to one. Then $\alpha \in (-1, 1)$ is identified by $g_{111}(\alpha) = (1-\alpha)(2r_{111} + 1 - \alpha) = 0$, where $g_{111}(\alpha)$ depends on data only via the median r_{111}. The Dhaene and Zhu (DZ) estimator $\hat{\alpha}_n$ then simply equals to $2\hat{r}_{n111} + 1$ and it was proved to be consistent and asymptotically normal (Dhaene and Zhu, 2009, Lemma 1).

To increase the precision of the estimation and possibly the robustness of the method, we propose to extend the DZ estimator by allowing for multiple differences. The full set of moment conditions in (3.8) can be written as

$$\boldsymbol{g}(\alpha) = \boldsymbol{0}, \tag{3.9}$$

where $\boldsymbol{g}(\alpha) = (g_{\boldsymbol{j}}(\alpha))_{\boldsymbol{j}\in\mathcal{J}}$ and a fixed finite set $\mathcal{J}$ contains all triplets $\boldsymbol{j} = (s, q, p)'$ that are considered in estimation. The DZ estimator corresponds then to the special case $\mathcal{J} = \{(1,1,1)'\}$. A set $\mathcal{J}$ with good robust properties will be constructed later in Section 3.3.

Since all equations in (3.9) have to be satisfied simultaneously, the parameter α is estimated by the generalized method of moments procedure:

$$\hat{\alpha}_n = \arg\min_{c \in (-1,1)} \boldsymbol{g}_n(c)' \boldsymbol{A}_n \boldsymbol{g}_n(c), \tag{3.10}$$

where $\boldsymbol{g}_n(c) = (g_{n\boldsymbol{j}}(c))_{\boldsymbol{j} \in \mathcal{J}}$ is the sample analog of $\boldsymbol{g}(\alpha)$ and corresponds to (3.8) with $r_{\boldsymbol{j}}$ being replaced by $\hat{r}_{n\boldsymbol{j}}$ defined in (3.6). The weighting matrix $\boldsymbol{A}_n$ has to be positive definite. Its simplest choice can be proportional to the number of observations available for the estimation of each moment equation: $\boldsymbol{A}_n = \boldsymbol{A} = \operatorname{diag}\{(T-p-q)/T\}$. The optimal choice of the GMM weighting matrix $\boldsymbol{A}$, which can be constructed only after an initial estimate of α is available, generally equals the inverse of the variance of $\boldsymbol{g}_n(\alpha)$. However, unreported simulation results show that the estimate of such optimal weighting matrix is inaccurate due to its size and complexity, and this results in a poor performance of (3.10).

The estimator defined in equation (3.10) will be referred to as the pairwise-difference DZ (PD-DZ) estimator. Its asymptotic distribution is derived in the following theorem using the following standard assumption on the weighting matrix $\boldsymbol{A}_n$. Note that we formulate the result for the simplicity of notation for a fixed T, but it can be derived also for $T \to \infty$.

A.4 Assume that $\boldsymbol{A}_n \xrightarrow{p} \boldsymbol{A}$ and $\boldsymbol{A}$ is positive definite.

Theorem 3. *Suppose that Assumptions A.1–A.4 hold. Let* $(1,1,1)' \in \mathcal{J}$ *and* $\boldsymbol{d} := \partial \boldsymbol{g}(\alpha)/\partial\alpha$, *where* α *represents the true parameter value. Then for a fixed* T *and* $n \to \infty$, $\hat{\alpha}_n$ *is consistent and asymptotically normal,*

$$\sqrt{n}(\hat{\alpha}_n - \alpha) \to \mathrm{N}(0, w), \tag{3.11}$$

where $w = (\boldsymbol{d}'\boldsymbol{A}\boldsymbol{d})^{-1}\boldsymbol{d}'\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{d}(\boldsymbol{d}'\boldsymbol{A}\boldsymbol{d})^{-1}$ *and* $\boldsymbol{V}$ *is defined in* (3.35) *in Lemma 2.*

3.3 Robustness properties

There are several measures of robustness. In the literature, they are usually divided in two main concepts, depending whether the sensitivity of an estimator to a finite (large) contamination of the data is studied (global or quantitative measures of robustness) or whether the sensitivity of an estimator to an infinitesimal contamination of the data is analysed (local or qualitative measures of robustness).

More formally, let $\mathcal{Z}$ be the set of all possible samples $Z^{\alpha} = \{z_{it}\}$ following model (3.1) and let $Z^{\epsilon,\zeta} = \{z^{\epsilon}_{it}\}$ be a contaminating sample following a fixed data-generating process, where the index ϵ of $Z^{\epsilon,\zeta}$ indicates the probability that an observation in $Z^{\epsilon,\zeta}$ is equal to $\zeta \neq 0$. The observed contaminated sample is $Z^{\alpha} + Z^{\epsilon,\zeta} = \{z_{it} + z^{\epsilon}_{it}\}$. Similarly to Dhaene and Zhu (2009), we consider the contamination by independent additive outliers and the contamination by patches of additive outliers. The former is defined as

$$Z^{1,\epsilon,\zeta} = \{z^{\epsilon}_{it}\}, \qquad P(z^{\epsilon}_{it} \neq 0) = \epsilon_1, \qquad P(z^{\epsilon}_{it} \leq u | z^{\epsilon}_{it} \neq 0) = G_{\zeta}(u), \tag{3.12}$$

where the distribution G_{ζ} with a parameter ζ could be left fully general but otherwise we consider a point mass distribution. The contamination by patches of k additive outliers is defined as

$$Z^{2,\epsilon,\zeta,k} = \{\zeta \cdot I(\nu^{\epsilon}_{it} = 1 \text{ or } \ldots \text{ or } \nu^{\epsilon}_{it-k+1} = 1)\}, \tag{3.13}$$

where ν^{ϵ}_{it} follows the Bernoulli distribution with the parameter ϵ_2 such that $(1-\epsilon_2)^k = \epsilon_1$. Additionally, a third contamination scheme $Z^{3,\epsilon,\zeta,k} = \{z^{\epsilon}_{it}\}$ is considered, where

$$z^{\epsilon}_{it} = \begin{cases} a_{it-l}(-1)^l & \text{if the smallest index } l \geq 0 \text{ with } \nu^{\epsilon}_{it-l} = 1 \text{ satisfies } l \leq k-1 \\ 0 & \text{otherwise,} \end{cases} \tag{3.14}$$

where $\Pr(a_{it-l} = \zeta) = 1/2$ and $\Pr(a_{it-l} = -\zeta) = 1/2$ and where ν^{ϵ}_{it} is defined as in $Z^{2,\epsilon,\zeta,k}$. Note that (3.13) and (3.14) are a special cases of a more general type of contamination $Z^{4,\epsilon,\zeta,k} = \{z^{\epsilon}_{it}\}$, where

$$z^{\epsilon}_{it} = \begin{cases} a_{it-l}(-\rho)^l & \text{if the smallest index } l \geq 0 \text{ with } \nu^{\epsilon}_{it-l} = 1 \text{ satisfies } l \leq k-1 \\ 0 & \text{otherwise,} \end{cases} \tag{3.15}$$

where $0 \leq \rho \leq 1$. We are not analysing this most general case and concentrate instead on the most extreme cases of $\rho = 1$ and $\rho = -1$ as they can arguably bias the estimate most.

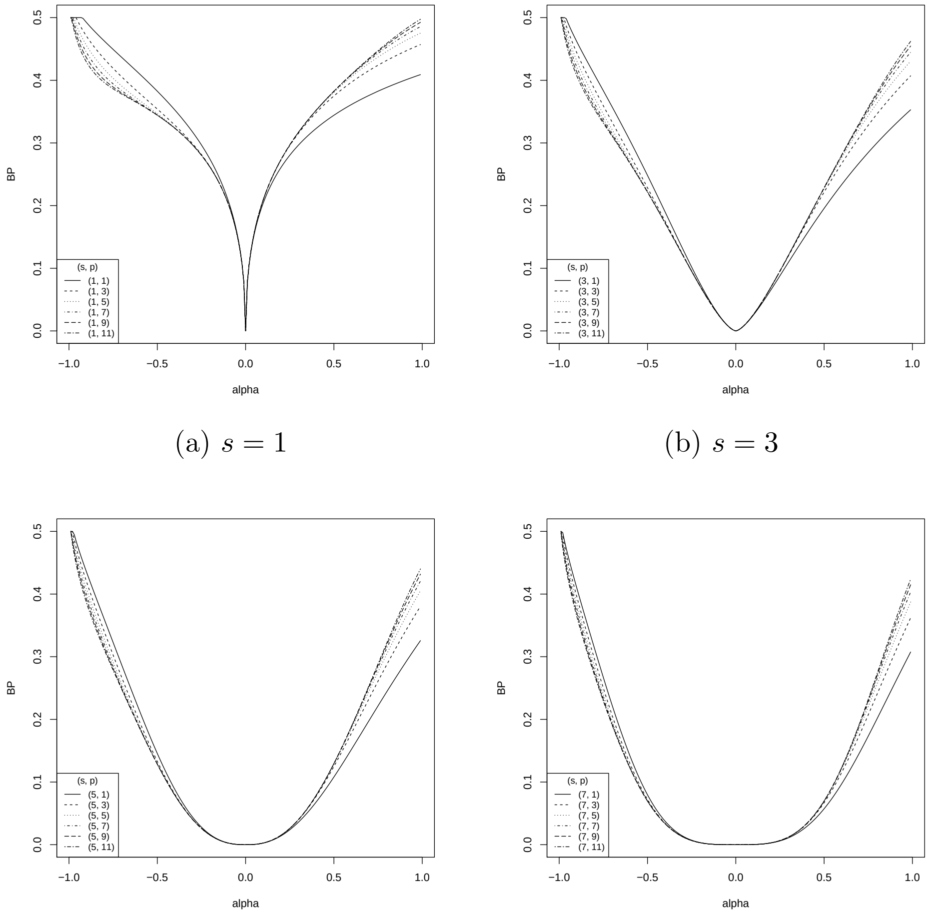
Before formally analyzing the robust properties of the PD-DZ estimator, note first that the median in (3.6) protects against outlying values of the ratio $\Delta^s y_{it}/\Delta^p y_{it-q}$ but not, in general, against large values of y_{it}. In other words, suppose that one observation y^{ζ}_{it-l} is contaminated, $y^{\zeta}_{it-l} = y_{it-l} + \zeta$, and $\zeta = +\infty$. Then the ratio $\Delta^s y_{it}/\Delta^p y_{it-q}$ becomes infinite in absolute value if $l = 0$, it becomes zero if $l = p + q$ or $l = q \neq s$, and it can become -1 if $l = s = q$. To protect against outliers, 0, -1, and $\pm\infty$ have to be outside or at the boundary of the $r_{\boldsymbol{j}}$ domain. As $\alpha \in (-1, 1)$, $r_{\boldsymbol{j}}$ in (3.7) attains negative, zero, and positive values for various values of α and thus 0 lies inside of its domain if $s \neq q$ or $s \neq p + q$. Such values should thus not be considered if a positive breakdown point is required. Assuming $s < p + q$ for simplicity, we impose the constraint $s \neq q$, which implies that $r_{\boldsymbol{j}}$ in (3.7) equals

$$r_{\boldsymbol{j}} = -\frac{(1-\alpha^s)(1-\alpha^p)}{2(1-\alpha^p)} = -\frac{1-\alpha^s}{2}. \tag{3.16}$$

Additionally, there should ideally be a one-to-one correspondence between the values of parameter α and the values of $r_{\boldsymbol{j}}$ in (3.16). Otherwise if $\alpha_1 \neq \alpha_2$ lead to the same value of $r_{\boldsymbol{j}}$, estimates could be easily biased from α_1 to α_2. For this reason, the power s of α in equation (3.16) has to be odd. Moreover, p has to be odd as well. As shown in Section 3.3.2, estimates of $\hat{r}_{n\boldsymbol{j}}$ given in Equation (3.6) become extremely sensitive to infinitesimal amount of contamination when they are computed by using also differences where p is even. Together with the constraint $s = q$ (i.e., assuming $s < p + q$ for simplicity), this results in the following choices of s, p, and q to be considered in the analysis of the breakdown point: $s = q$ and s and p are odd, that is, $\mathcal{J}_o = \{(s, q, p)' : s = q; s, p \text{ are odd}\}$. Finally, note that for simplicity the robustness of the proposed estimator is studied in a setting where errors are homoskedastic, $\sigma_i^2 = \sigma^2$ for all $i = 1, \cdots, n$.

This section is organized as follows. First, the breakdown point of the estimates $\hat{r}_{n\boldsymbol{j}}$ for $\boldsymbol{j} = (s, s, p)' \in \mathcal{J}_o$ is analyzed (Section 3.3.1). Next, the sensitivity of the estimates $\hat{r}_{n\boldsymbol{j}}$ to infinitesimal amount of contamination is studied (Section 3.3.2). Finally these results are extended to the GMM estimator in (3.10) (Section 3.3.3).

3.3.1 Asymptotic Breakdown point of the individual estimator $\hat{r}_{n\boldsymbol{j}}$

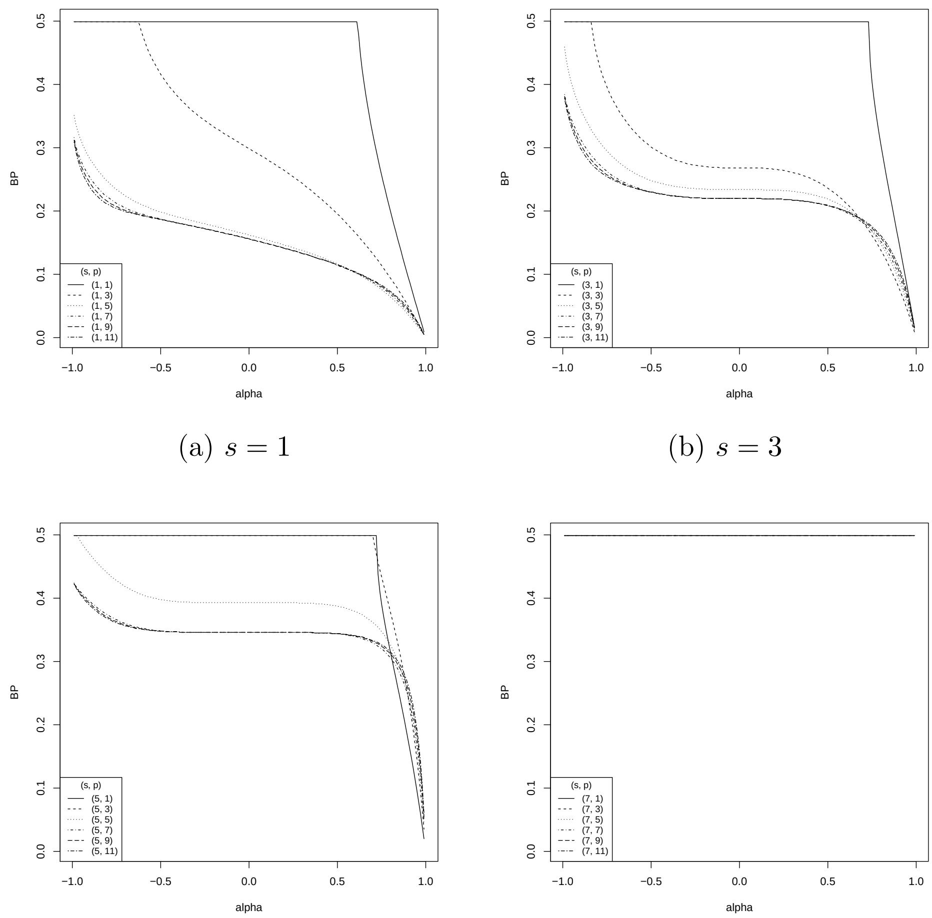


(a) $s = 1$ (b) $s = 3$

(c) $s = 5$ (d) $s = 7$

Figure 3.1: Breakdown point of $\hat{r}_{\boldsymbol{j}}$, $\boldsymbol{j} = (s, s, p)' \in \mathcal{J}_o$, under contamination $Z^{1,\epsilon,\zeta}$ by independent additive outliers.

One of the traditional measures of the global robustness of an estimator is the breakdown point. It can be defined as the smallest fraction of the data that can be changed in such a way that the estimator will not reflect any information concerning the remaining (non-contaminated) observations. Let $\mathcal{T}$ denote a generic estimator of an unknown parameter θ. Similarly to Dhaene and Zhu (2009), the breakdown point will be defined using the limit of $\mathcal{T}(Z^{\theta}_{nT})$ of θ evaluated at samples $Z^{\theta}_{nT} + Z^{\mathfrak{I},\epsilon,\zeta}_{nT}$ of size (n, T), where the samples $Z^{\theta}_{nT} = \{z_{it}\}^{n,\ T}_{i=1,t=1}$ follow model (3.1), where $Z^{\mathfrak{I},\epsilon,\zeta}_{nT}$ represents contaminating samples following one of the data-generating process $\mathfrak{I} = 1, 2, 3$ in (3.12)–(3.14) fully determined by a parameter ζ and the probability ϵ of



(a) $s = 1$

(b) $s = 3$

(c) $s = 5$

(d) $s = 7$

Figure 3.2: Breakdown point of $\hat{r}_{\boldsymbol{j}}$, $\boldsymbol{j} = (s, s, p)' \in \mathcal{J}_o$, under contamination $Z^{2,\epsilon,\zeta,k}$ by patch additive outliers, length of the path $k = 6$.

data contamination, and where $Z^{\theta}_{nT} + Z^{\Im,\epsilon,\zeta}_{nT} = \{z_{it} + z^{\epsilon}_{it}\}^{n}_{i=1},{}^{T}_{t=1}$. Denoting the limit $\mathcal{T}(Z^{\theta} + Z^{\Im,\epsilon,\zeta}) = \text{plim}_{n\to\infty} \mathcal{T}(Z^{\theta}_{nT} + Z^{\Im,\epsilon,\zeta}_{nT})$, the breakdown point ϵ^* of an estimator $\mathcal{T}$ of θ to a constant $c \neq \theta$, can be defined as

$$\epsilon^*\big(\mathcal{T}(Z^{\theta}+Z^{\Im,\epsilon,\zeta}); c\big) := \inf_{\epsilon\geq 0} \left\{ \epsilon \,\middle|\, \sup_{\zeta} \mathcal{T}(Z^{\theta} + Z^{\Im,\epsilon,\zeta}) = \inf_{\zeta} \mathcal{T}(Z^{\theta} + Z^{\Im,\epsilon,\zeta}) = c \right\}. \tag{3.17}$$

The breakdown point thus characterizes the smallest fraction of contamination that will drive the estimator $\mathcal{T}(Z^{\theta} + Z^{\Im,\epsilon,\zeta})$ to c, where the data contamination $Z^{\Im,\epsilon,\zeta}$, $\Im = 1, 2, 3$, is allowed to be defined by any value of parameter ζ, but it follows one of the proposed data-generating process (3.12)–(3.14).

More than one definition of breakdown point has been proposed in the literature (see Davies and Gather, 2005, and references therein). The definition in (3.17) was originally suggested by Genton and Lucas (2003) and it has been criticized mainly for two reasons, see Davies and Gather (2005). Firstly, it depends on the type of contamination, in this case $Z^{\Im,\epsilon,\zeta}$. It has been argued that, contrary to (3.17), the breakdown point of an estimator should be derived independently of the process underling the contamination of the data. Secondly, for some data contamination it may lack of a finite sample analogue as its "continuous formulations do not approximate the discrete world of statistics" (Davies and Gather, 2005, rejoinder, second paragraph).

Despite these criticisms, the definition in (3.17) continues to be used, especially in some nonlinear models and models under dependence, see among others Azzalini and Genton (2008), Dhaene and Zhu (2009), Muler et al. (2009), and Čížek (2012). The reason is that there is no exhaustive definition in the literature, especially for models like the one in (3.1). This point can be illustrated by means of the so-called "classical definition" as presented in Davies and Gather (2005, Eq. (2.4)). This states that the asymptotic breakdown point ϵ^* of an estimator $\mathcal{T}$ of θ is defined as

$$\epsilon^*(\mathcal{T}; Z^{\theta}, D) := \inf_{\epsilon\geq 0} \left\{ \epsilon \,\middle|\, \sup_{Z^{\epsilon,\zeta}\in\mathcal{Z}^{\epsilon}} D\Big(\mathcal{T}(Z^{\theta}), \mathcal{T}(Z^{\theta} + Z^{\epsilon,\zeta})\Big) = \infty \right\}, \tag{3.18}$$

where $D(\cdot,\cdot)$ is an appropriately chosen metric and $\mathcal{Z}^{\epsilon}$ is the set of all possible data contaminations — not just those proposed in (3.12)–(3.14).

Cases where the support of the estimator is bounded, $\mathcal{T} \in (a, b)$, $a, b \in \mathbb{R}$, are easily accommodated in (3.18) by choosing $D\colon [a, b] \to \mathbb{R}$ such that $D\Big(\mathcal{T}(Z^\theta), \mathcal{T}(Z^\theta + Z^{\epsilon,\zeta})\Big) \to \infty \Leftrightarrow \mathcal{T}(Z^\theta + Z^{\epsilon,\zeta}) \to c$, where c is one of the parameter space boundaries, $c = a$ or $c = b$.

The limitation of the classical definition (3.18) is its inapplicability to cases where the estimator breaks down to a point within the domain of the estimator. For instance, consider the ordinary least squares (OLS) estimator of an autoregressive time series model of order 1,

$$y_t = \phi y_{t-1} + \varepsilon_t \qquad (t = 2, \cdots, T),$$

with bounded parameter space $\phi \in (-1, 1)$ and with error terms independent of each others. Even in the case where only one observation is contaminated $y_1^\zeta = y_1 + \zeta$, the OLS estimate based on the contaminated sample $(y_1^\zeta, y_2, \cdots, y_T)'$ will $\hat{\phi}^{(OLS)} \to 0$ as $\zeta \to \infty$. However, the fact that the estimator is totally determined by the contamination is not captured as break down behavior under the classical definition.

Similarly, let us consider the bias toward zero caused by independent additive outliers in model (3.1) to the estimator (3.10). Despite the fact that the estimator "collapses" to zero and becomes totally uninformative because of the contamination (see Theorem 4), the estimator is not said to have broken down according to (3.18).

To the best of our knowledge, no definition of breakdown point proposed in the literature is free from criticisms under any possible type of data contamination and any model specification. In this paper, we use (3.17) restricting the analysis to three of the most representative data contaminations (3.12)–(3.14).

As a final remark, let us show that the two definitions (3.17) and (3.18) are mutually linked under the three data contaminations (3.12)–(3.14). As mentioned before, this is certainly the case for breakdown of $\hat{\alpha}^{PD-DZ}$ to the boundaries of the parameters space ± 1 (see Theorems 5 and 6). Similarly, the "collapsing" behavior of the estimator to $c = 0$ due to independent additive outliers (Theorem 4) has the same properties as breakdown to the boundary of the parameter space under the classical definition: if the amount of contamination increases above a certain level $\epsilon^*(\hat{\alpha}^{PD-DZ}(Z^\alpha + Z^{1,\epsilon,\zeta}))$, then $\hat{\alpha}^{PD-DZ}$ can be made arbitrarily close to zero. Hence, $c = 0$ acts as boundary of the parameter space of $|\alpha|$. Note: the fact that the estimator collapses to zero rather than to another point on the support of α is guaranteed by Theorem 1 in Dhaene and Zhu (2009), which is applicable to the general PD-DZ estimator, and which states that independent additive outliers (i) the bias of $\hat{\alpha}^{PD-DZ}$ due to contamination has opposite sign to α, (ii) the bias is bounded in absolute value between zero and $|\alpha|$, and (iii) the bias in absolute value is increasing with $|\alpha|$ on $(0, 1)$.

The following results concerning the breakdown point of $\hat{r}_{\boldsymbol{j}}$ are generalizations of Theorems 5 and 8 of Dhaene and Zhu (2009). The first one characterizes the breakdown behavior under the independent-outlier contamination, whereas the next two do this under the contamination by the patches of outliers.

Theorem 4. *Consider the independent additive outlier contamination* $Z^{1,\epsilon,\zeta}$ *occurring with probability* ϵ_1*, where* $0 < \epsilon_1 < 1$*. Then* $\hat{r}_{\boldsymbol{j}}$*,* $\boldsymbol{j} = (s, s, p)' \in \mathcal{J}_o$*, breaks down to* $c = -1/2$ *(irrespective of* ζ*) if* ϵ_1 *is smaller than the smallest positive root of the equation*

$$\frac{100}{\pi}\frac{1-\epsilon}{25-17\epsilon}\left(\frac{1-\epsilon}{\epsilon}\right)^2 \arctan\frac{|\alpha^s/2|}{\sqrt{4(1-\alpha^s)/(1-\alpha^p)-(1-\alpha^s)^2}} = 1. \tag{3.19}$$

Theorem 5. *Consider the patched additive outlier contamination* $Z^{2,\epsilon,\zeta,k}$ *occurring with probability* ϵ*, where* $0 < \epsilon < 1$*,* $|\zeta| \to \infty$*, and the length of the patches* $k \geq 2$*. Then* $\hat{r}_{\boldsymbol{j}}$*,* $\boldsymbol{j} = (s, s, p)' \in \mathcal{J}_o$*, breaks down to* $c = 0$ *as* $|\zeta| \to \infty$ *if and only if*

$$2\frac{1-\mathfrak{p}_B-\mathfrak{p}_C-\mathfrak{p}_D}{\mathfrak{p}_C-\mathfrak{p}_D}\frac{1}{\pi}\arctan\left(\sqrt{\frac{1-\alpha^p}{1-\alpha^s-(1-\alpha^s)^2(1-\alpha^p)/4}}\frac{1-\alpha^s}{2}\right) \leq 1, \tag{3.20}$$

where $\mathfrak{p}_B$*,* $\mathfrak{p}_C$*, and* $\mathfrak{p}_D$ *are defined in* (3.67)–(3.69)*, respectively.*

Theorem 6. *Consider the patched additive outlier contamination* $Z^{3,\epsilon,\zeta,k}$ *occurring with probability* ϵ*, where* $0 < \epsilon < 1$*,* $|\zeta| \to \infty$*, and the length of the patches* $k \geq 2$*. Then* $\hat{r}_{\boldsymbol{j}}$*,* $\boldsymbol{j} = (s, s, p)' \in \mathcal{J}_o$*, breaks down to* $c = -1$ *as*

$|\zeta| \to \infty$ if and only if

$$2\frac{\mathfrak{p}_A}{\mathfrak{p}_C + \mathfrak{p}_E + \mathfrak{p}_F + \mathfrak{p}_H - (\mathfrak{p}_D + \mathfrak{p}_G + \mathfrak{p}_I)}\frac{1}{\pi} \times \arctan\left(-\sqrt{\frac{1-\alpha^p}{1-\alpha^s-(1-\alpha^s)^2(1-\alpha^p)/4}}\frac{1+\alpha^s}{2}\right) \leq 1, \quad (3.21)$$

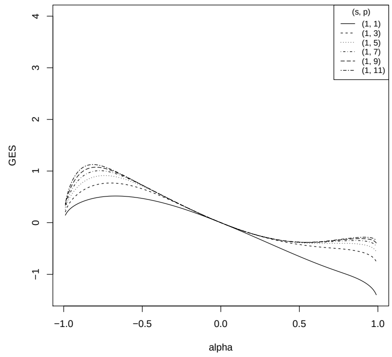
where $\mathfrak{p}_j$, $j = \{C, D, E, F, G, H, I, A\}$, are defined in Equations (3.88)–(3.96) *in Appendix 3.A.2.*

The lower bounds for the breakdown points of $\hat{r}_j$ derived in Theorems 4–6 are displayed in Figures 3.1 and 3.2, respectively. In the case of independent additive outliers, median ratios based on $s = 1$ have generally rather similar breakdown points and are superior or comparable to those using $s = 3$ (or higher) for any value of α. In other words, the combination of $s = 1$ and p odd yields the highest breakdown points against independent additive outliers. In the case of patches of additive outliers $Z^{2,\epsilon,\zeta,k}$, the opposite conclusion holds: the breakdown point increases with an increasing s and deteriorates with an increasing p.

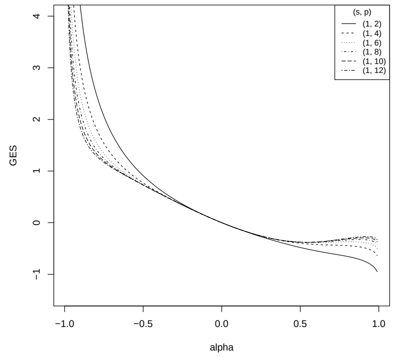
3.3.2 Qualitative measures of robustness

Qualitative or local measures of robustness express the sensitivity of an estimator to an infinitesimal (small) amount of data contamination. One of the traditional measure of local robustness is the influence function. This is defined as follows. Let $\mathcal{T}(Z^\theta_{nT} + Z^{\mathfrak{I},\epsilon,\zeta}_{nT})$ denote a generic estimator of an unknown parameter θ based on a contaminated sample $Z^\theta_{nT} + Z^{\mathfrak{I},\epsilon,\zeta}_{nT} = \{z_{it} + z^{\epsilon}_{it}\}^{n,\ T}_{i=1,t=1}$, where Z^θ_{nT} and $Z^{\mathfrak{I},\epsilon,\zeta}_{nT}$ have been defined at the beginning of Section 3.3. Let $\mathcal{T}(Z^\theta + Z^{\mathfrak{I},\epsilon,\zeta})$ be its probability limit when T is fixed and $n \to \infty$. Note that $\mathcal{T}$ depends of the unknown parameter θ, of the fraction of data contamination ϵ, and of the non-zero value of the outlier ζ. Assume $\mathcal{T}$ is consistent under non-contaminated data, that is $\mathcal{T}(Z^\theta) = \theta$. The influence function (IF) is defined as

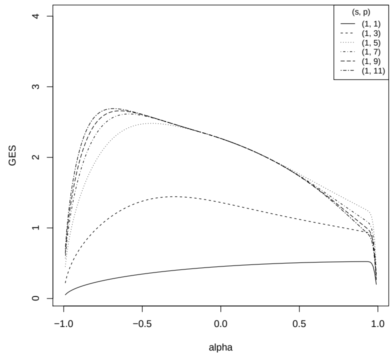
$$\mathrm{IF}\left(\mathcal{T}; Z^\theta, Z^{\mathfrak{I},\epsilon,\zeta}\right) := \lim_{\epsilon\to 0}\frac{\mathcal{T}(Z^\theta + Z^{\mathfrak{I},\epsilon,\zeta}) - \theta}{\epsilon} = \left.\frac{\partial\,\mathrm{bias}(\mathcal{T}; Z^\theta, Z^{\mathfrak{I},\epsilon,\zeta})}{\partial\epsilon}\right|_{\epsilon=0}, \quad (3.22)$$



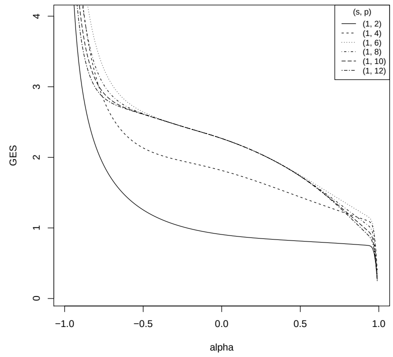
(a) p odd



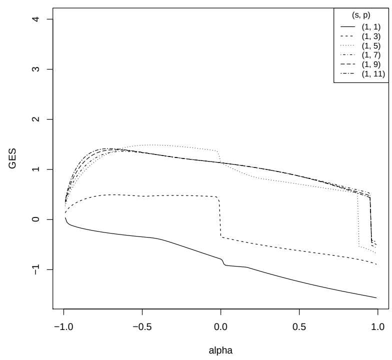
(b) p even



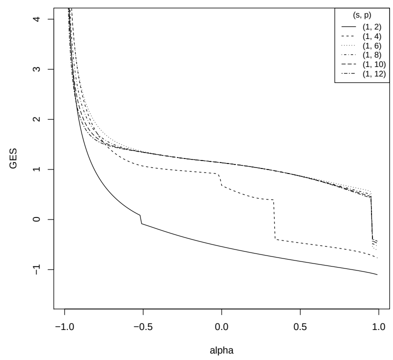
(c) p odd



(d) p even

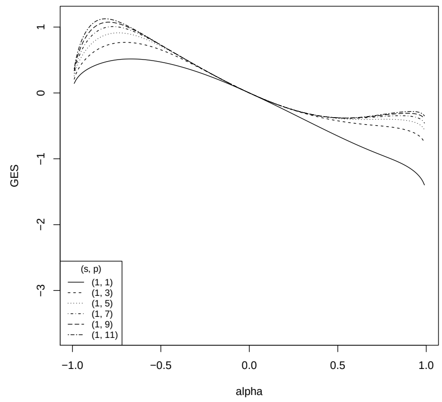


(e) p odd

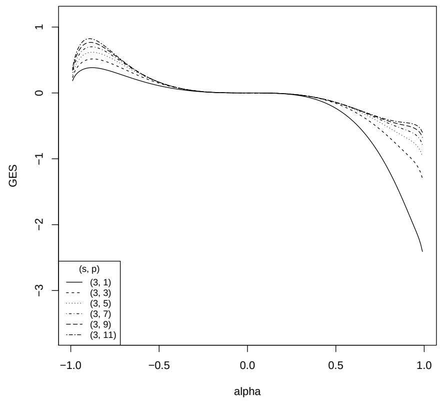


(f) p even

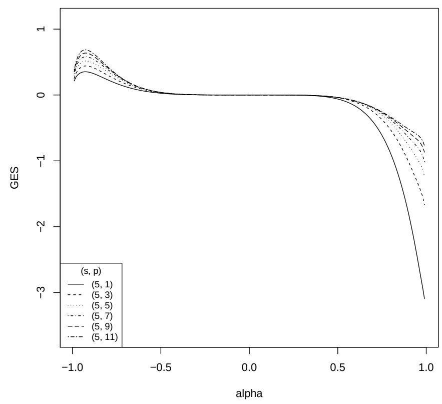
Figure 3.3: Gross-error sensitivity of $\hat{r}_{\boldsymbol{j}}$, for $s = 1$, under contamination $Z^{1,\epsilon,\zeta}$ (first row), $Z^{2,\epsilon,\zeta,k}$ with $k = 6$ (second row), and $Z^{3,\epsilon,\zeta,k}$ with $k = 6$ (third row).



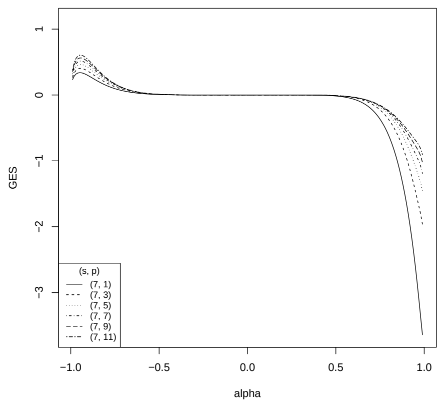
(a) $s = 1$



(b) $s = 3$

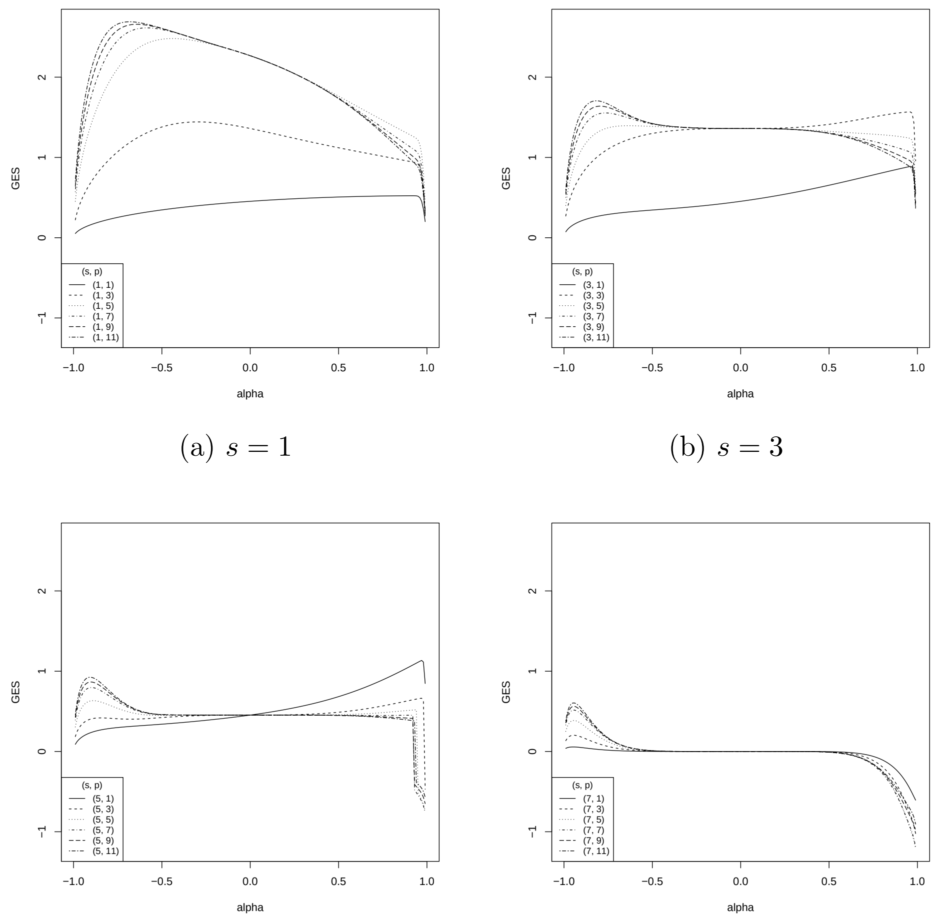


(c) $s = 5$



(d) $s = 7$

Figure 3.4: Gross-error sensitivity of $\hat{r}_{\boldsymbol{j}}$, $\boldsymbol{j} = (s, s, p)' \in \mathcal{J}_o$, under contamination $Z^{1,\epsilon,\zeta}$ by independent additive outliers.



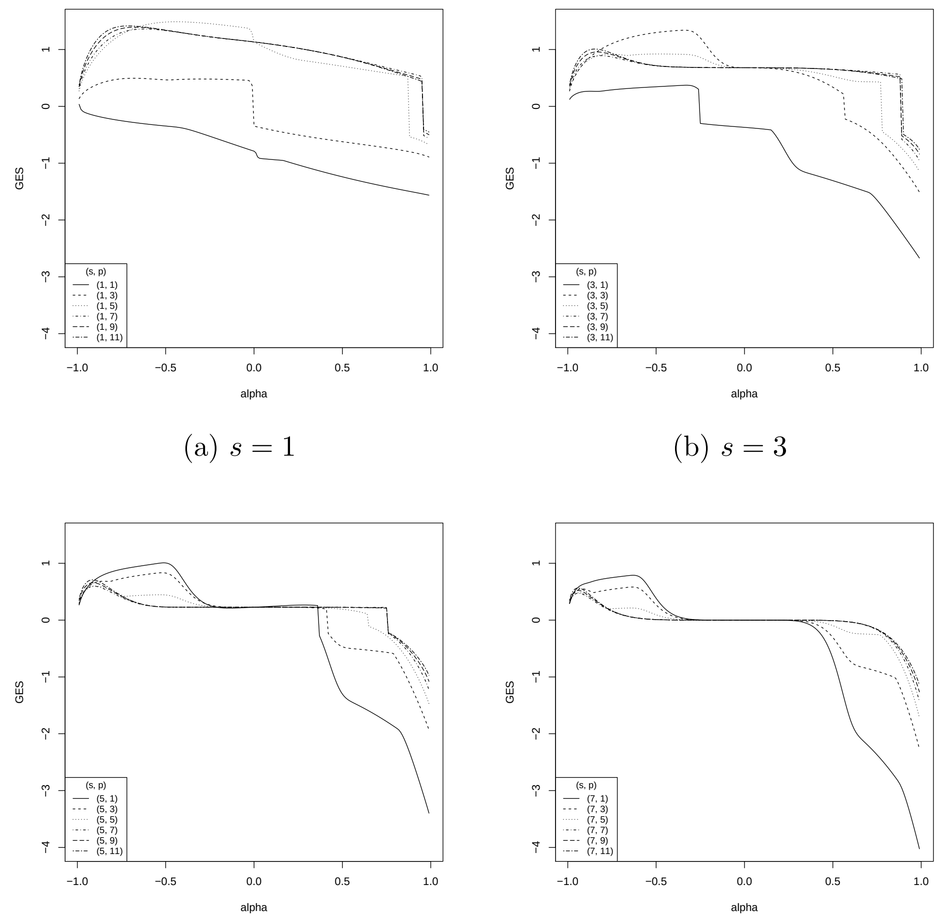
(a) $s = 1$

(b) $s = 3$

(c) $s = 5$

(d) $s = 7$

Figure 3.5: Gross-error sensitivity of $\hat{r}_{\boldsymbol{j}}$, $\boldsymbol{j} = (s, s, p)' \in \mathcal{J}_o$, under contamination $Z^{2,\epsilon,\zeta,k}$ by patch additive outliers, length of the path $k = 6$.



(a) $s = 1$

(b) $s = 3$

(c) $s = 5$

(d) $s = 7$

Figure 3.6: Gross-error sensitivity of $\hat{r}_{\boldsymbol{j}}$, $\boldsymbol{j} = (s, s, p)' \in \mathcal{J}_o$, under contamination $Z^{3,\epsilon,\zeta,k}$ by patch additive outliers, length of the path $k = 6$.

where the equality follows by the definition of asymptotic bias of $\mathcal{T}(Z^\theta + Z^{\mathfrak{J},\epsilon,\zeta})$ due to the data contamination

$$\text{bias}\left(\mathcal{T}; Z^\theta, Z^{\mathfrak{J},\epsilon,\zeta}\right) := \mathcal{T}(Z^\theta + Z^{\mathfrak{J},\epsilon,\zeta}) - \theta. \tag{3.23}$$

Building on Dhaene and Zhu (2009, Theorems 2 and 7), we derive the following results for the PD-DZ estimator under the three considered data contaminations:

Theorem 7. *Under independent additive contamination $Z^{1,\epsilon,\zeta}$ with point-mass distribution at $\zeta \neq 0$,*

$$\begin{aligned}\mathrm{IF}(\hat{r}_{\boldsymbol{j}}; Z^\alpha, Z^{1,\epsilon,\zeta}) &= -\pi\sqrt{\frac{1-\alpha^s}{1-\alpha^p} - \frac{1}{4}(1-\alpha^s)^2} \\ &\times \left[\Phi\left(\frac{\zeta(1+\alpha^s)/2}{\sqrt{2\frac{\sigma_\epsilon^2}{1-\alpha^s}\left(1-\alpha^s - \frac{(1-\alpha^s)^2}{4}(1-\alpha^p)\right)}}\right) - \right. \\ &\left.\Phi\left(\frac{\zeta(1-\alpha^s)/2}{\sqrt{2\frac{\sigma_\epsilon^2}{1-\alpha^s}\left(1-\alpha^s - \frac{(1-\alpha^s)^2}{4}(1-\alpha^p)\right)}}\right)\right] \\ &\times\left[\Phi\left(\frac{\zeta}{\sqrt{2\sigma_\epsilon^2\frac{1-\alpha^p}{1-\alpha^s}}}\right) - \Phi\left(-\frac{\zeta}{\sqrt{2\sigma_\epsilon^2\frac{1-\alpha^p}{1-\alpha^s}}}\right)\right].\end{aligned} \tag{3.24}$$

Theorem 8. *Consider the patched additive outlier contamination $Z^{2,\epsilon,\zeta,k}$ with point-mass distribution at $\zeta \neq 0$ and patch length $k \geq 2$. Then*

$$\begin{aligned}\mathrm{IF}(\hat{r}_{\boldsymbol{j}}; Z^\alpha, Z^{2,\epsilon,\zeta,k}) &= -\frac{\pi}{k}\sqrt{\frac{1-\alpha^s}{1-\alpha^p} - \frac{(1-\alpha^s)^2}{4}} \\ &\times\left[\mathfrak{p}_C'(0)\left(C(r_{\boldsymbol{j}};\zeta,0) - \frac{1}{2}\right) + \mathfrak{p}_D'(0)\left(D(r_{\boldsymbol{j}};\zeta,0) - \frac{1}{2}\right)\right],\end{aligned} \tag{3.25}$$

where $\mathfrak{p}_C'(0)$, $\mathfrak{p}_D'(0)$, $C(r_{\boldsymbol{j}};\zeta,0)$, and $D(r_{\boldsymbol{j}};\zeta,0)$ are defined in (3.80), (3.81), (3.84), *and* (3.85)*, respectively.*

Theorem 9. *Consider the patched additive outlier contamination $Z^{3,\epsilon,\zeta,k}$ with point-mass distribution at $\zeta \neq 0$ and patch length $k \geq 2$. Denote*

$\mathcal{L}(1/2) := L(r_{\boldsymbol{j}};\zeta,0) - 1/2$, for $\mathcal{L} \in \{\mathcal{C},\mathcal{D},\mathcal{E},\mathcal{G},\mathcal{I}\}$. Then

$$\begin{aligned}\mathrm{IF}(\hat{r}_{\boldsymbol{j}};Z^{\alpha},Z^{\mathfrak{Z},\epsilon,\zeta,k}) &= -\frac{\pi}{k}\sqrt{\frac{1-\alpha^s}{1-\alpha^p} - \frac{(1-\alpha^s)^2}{4}} \\ &\times \left[\mathfrak{p}'_C\mathcal{C}\left(\frac{1}{2}\right) + \mathfrak{p}'_D\mathcal{D}\left(\frac{1}{2}\right) + \mathfrak{p}'_E\mathcal{E}\left(\frac{1}{2}\right) + \mathfrak{p}'_G\mathcal{G}\left(\frac{1}{2}\right) + \mathfrak{p}'_I\mathcal{I}\left(\frac{1}{2}\right)\right]\end{aligned} \tag{3.26}$$

where $\mathfrak{p}'_L$ and L, $L \in \{C,D,E,G,I\}$, are defined in Equations (3.106), (3.107), (3.108), (3.110), (3.112), *and* (3.115)–(3.119), *in Appendix 3.A.2.*

The influence functions reported in Theorems 7–9 depend on the value of parameter ζ. Hence, one may derive which value of ζ yields the large sensitivity of an estimator, that is the largest value of IF function in absolute value. This is the definition of gross error sensitivity (GES). More formally, this is defined as

$$\mathrm{GES}\left(\mathcal{T};Z^{\theta},Z^{\mathfrak{Z},\epsilon,**}\right) := \mathrm{IF}\left(\mathcal{T};Z^{\theta},Z^{\mathfrak{Z},\epsilon,\zeta^{**}}\right), \tag{3.27}$$

where

$$\zeta^{**} := \arg\sup_{\zeta}\left|\mathrm{IF}\left(\mathcal{T};Z^{\theta},Z^{\mathfrak{Z},\epsilon,\zeta}\right)\right|. \tag{3.28}$$

Given the results in Theorems 7–9, we compute numerically the GES curves for each $\hat{r}_{\boldsymbol{j}}$. The variance σ_{ε}^2 is set equal to one without loss of generality. We first focus on $\boldsymbol{j} = (s,s,p)'$ where p can take both even and odd values. As shown in Figure 3.3, GES curves are bounded only if p even is used. The estimator $\hat{r}_{\boldsymbol{j}}(Z^{\alpha} + Z^{\mathfrak{Z},\epsilon,\zeta})$ becomes extremely sensitive to infinitesimal amount of contamination when even values of p are used. Such values of p should not be considered if a bounded influence function is required. Next, we consider $\boldsymbol{j} = (s,s,p)' \in \mathcal{J}_o$, see Figures 3.4–3.6. All curves are bounded in the interval $\alpha \in (-1,1)$ and display generally more sensitivity for $|\alpha|$ close to one than for values of the autoregressive parameter around zero. Finally, note that median ratios become insensitive when the order of the s-difference is larger than the length of the patch, that is for $s > k$.

3.3.3 Robust properties of the GMM estimator $\hat{\alpha}_n$

Given the results of the previous sections, we will now analyze the robust properties of the general GMM estimator $\hat{\alpha}_n$ defined in (3.10) and based on moment equations (3.9) for $\boldsymbol{j} = (s,s,p)' \in \mathcal{J}_o$. For the sake of simplicity, we assume now that the weighting matrix of the PD-DZ estimator (3.10) is sample independent, $\boldsymbol{A}_n = \boldsymbol{A}$, for example, $\boldsymbol{A} = \mathrm{diag}\{(T-s-p)/T\}$ as discussed in Section 3.2. Otherwise, the breakdown properties of a particular definition of the data-dependent $\boldsymbol{A}_n$ would have to be known.

Theorem 10. *Consider a particular additive outlier contamination $Z^{\mathfrak{I},\epsilon,\zeta}$ occurring with probability ϵ, where $0 < \epsilon < 1$. Further, let all $\hat{r}_{\boldsymbol{j}}(Z^\alpha + Z^{\mathfrak{I},\epsilon,\zeta})$, $\boldsymbol{j} = (s,s,p)' \in \mathcal{J} \subseteq \mathcal{J}_o$, break down to a finite constant c_b with breakdown points $\epsilon^*(\hat{r}_{\boldsymbol{j}}(Z^\alpha + Z^{\mathfrak{I},\epsilon,\zeta}); c_b)$. Finally, assume that $\boldsymbol{A}_n = \boldsymbol{A}$ is a positive definite diagonal matrix. Then the breakdown point of the GMM estimator $\hat{\alpha}(Z^\alpha + Z^{\mathfrak{I},\epsilon,\zeta})$ using moment conditions indexed by $\mathcal{J}$ is greater or equal to their maximum:*

$$\epsilon^*\Big(\hat{\alpha}(Z^\alpha + Z^{\mathfrak{I},\epsilon,\zeta}); c_b\Big) \geq \max_{\boldsymbol{j}\in\mathcal{J}} \epsilon^*\Big(\hat{r}_{\boldsymbol{j}}(Z^\alpha + Z^{\mathfrak{I},\epsilon,\zeta}); c_b\Big). \tag{3.29}$$

Theorem 11. *Consider a particular additive outlier contamination $Z^{\mathfrak{I},\epsilon,\zeta}$ occurring with probability ϵ, where $0 < \epsilon < 1$. Further, let all $\hat{r}_{\boldsymbol{j}}$, $\boldsymbol{j} = (s,s,p)' \in \mathcal{J} \subseteq \mathcal{J}_o$. Finally, assume that $\boldsymbol{A}_n = \boldsymbol{A}$ is a positive definite diagonal matrix. Then the influence function of the GMM estimator $\hat{\alpha}$ using moment conditions indexed by $\mathcal{J}$ is given by*

$$\mathrm{IF}(\hat{\alpha}; Z^\alpha, Z^{\mathfrak{I},\epsilon,\zeta}) = -(\boldsymbol{d}'\boldsymbol{A}\boldsymbol{d})^{-1}\boldsymbol{d}'\boldsymbol{A}\boldsymbol{\psi}, \tag{3.30}$$

where $\boldsymbol{d}$ is defined in Theorem 3 and $\boldsymbol{\psi}$ is the $\#\mathcal{J}_o \times 1$ vector of the influence function of each single $\hat{r}_{\boldsymbol{j}}$, $\boldsymbol{\psi} := \big(\mathrm{IF}(\hat{r}_{\boldsymbol{j}}; Z^\alpha, Z^{\mathfrak{I},\epsilon,\zeta})\big)_{\boldsymbol{j}\in\mathcal{J}_o}$.

Hence, the proposed PD-DZ estimator does not decrease, and in general, increases the breakdown point with respect to the original DZ estimator ($(1,1,1)' \in \mathcal{J}_o$). A final remark concerns the bias of the discussed estimators: although the breakdown point of the PD-DZ estimator equals to the maximum of the breakdown points of individual median estimators, the bias of PD-DZ due to contamination is not equal to the smallest bias achieved by one of the median estimators $\hat{r}_{n\boldsymbol{j}}$, but rather to a linear combination of those biases. Therefore, it is not wise to add moment conditions for triplets $\boldsymbol{j}$ that are not robust: while this would likely increase the precision of estimation, it could induce large biases under data contamination.

Table 3.1: Monte Carlo simulations: biases and RMSE for all estimators in model with $\varepsilon_{it} \sim \mathrm{N}(0,1)$ and $\eta_i \sim N(0,1)$ under different sample sizes ($T = 12$ and n increasing).

		Bias				RMSE			
α	n	25	50	100	200	25	50	100	200
0.1	AB	-0.0508	-0.0291	-0.0127	-0.0075	0.090	0.064	0.042	0.031
	BB	-0.0581	-0.0345	-0.0130	-0.0080	0.099	0.069	0.046	0.034
	DZ	0.0041	-0.0002	0.0048	0.0009	0.160	0.117	0.081	0.061
	PD-DZ ($s = 1$)	0.0057	0.0008	0.0015	0.0003	0.125	0.091	0.061	0.046
	PD-DZ ($s \geq 1$)	0.0073	0.0015	0.0018	0.0005	0.128	0.092	0.062	0.046
0.5	AB	-0.1015	-0.0579	-0.0344	-0.0167	0.137	0.088	0.060	0.039
	BB	-0.0673	-0.0434	-0.0250	-0.0120	0.117	0.082	0.057	0.041
	DZ	-0.0075	0.0019	0.0014	0.0023	0.181	0.125	0.090	0.065
	PD-DZ ($s = 1$)	-0.0099	-0.0006	-0.0005	0.0021	0.116	0.084	0.060	0.043
	PD-DZ ($s \geq 1$)	-0.0071	-0.0003	-0.0009	0.0025	0.117	0.084	0.059	0.042
0.9	AB	-0.2898	-0.2548	-0.2033	-0.1318	0.316	0.281	0.229	0.155
	BB	0.0346	0.0342	0.0311	0.0256	0.059	0.054	0.049	0.044
	DZ	0.0119	0.0033	-0.0009	-0.0033	0.192	0.139	0.096	0.070
	PD-DZ ($s = 1$)	-0.0099	-0.0060	-0.0011	-0.0014	0.097	0.078	0.057	0.042
	PD-DZ ($s \geq 1$)	-0.0137	-0.0054	-0.0019	-0.0023	0.090	0.066	0.046	0.032

Table 3.2: Monte Carlo simulations: biases and RMSE for all estimators in model with $\varepsilon_{it} \sim \mathrm{N}(0,1)$ and $\eta_i \sim N(0,1)$ under different sample sizes ($n = 100$ and T increasing).

		Bias			RMSE		
α	T	6	12	24	6	12	24
0.1	AB	-0.0174	-0.0127	-0.0124	0.083	0.042	0.027
	BB	-0.0032	-0.0130	-0.0404	0.079	0.046	0.049
	DZ	0.0013	0.0048	-0.0041	0.132	0.081	0.056
	PD-DZ ($s = 1$)	-0.0004	0.0015	-0.0015	0.121	0.061	0.037
	PD-DZ ($s \geq 1$)	-0.0004	0.0018	-0.0014	0.122	0.062	0.038
0.5	AB	-0.0495	-0.0344	-0.0227	0.127	0.060	0.035
	BB	-0.0022	-0.0250	-0.0651	0.096	0.057	0.073
	DZ	-0.0011	0.0014	0.0028	0.146	0.090	0.061
	PD-DZ ($s = 1$)	-0.0001	-0.0005	0.0005	0.128	0.060	0.033
	PD-DZ ($s \geq 1$)	-0.0013	-0.0009	0.0008	0.124	0.059	0.035
0.9	AB	-0.4499	-0.2033	-0.0863	0.552	0.229	0.094
	BB	0.0464	0.0311	-0.0011	0.080	0.049	0.030
	DZ	-0.0040	-0.0009	0.0014	0.153	0.096	0.065
	PD-DZ ($s = 1$)	-0.0200	-0.0011	0.0004	0.110	0.057	0.030
	PD-DZ ($s \geq 1$)	-0.0182	-0.0019	-0.0006	0.099	0.046	0.022

Table 3.3: Monte Carlo simulations: RMSE for all estimators in models with $\varepsilon_{it} \sim$ N$(0,1)$ and $\eta_i \sim N(0,\sigma_\eta^2)$, $\sigma_\eta^2 \in \{1/4,1,4\}$, under different sample sizes.

	(n,T)	$(100,6)$			$(50,12)$			$(25,24)$		
α	$\sigma_\eta^2/\sigma_\varepsilon^2$	1/4	1	4	1/4	1	4	1/4	1	4
0.1	AB	0.071	0.081	0.091	0.060	0.063	0.066	0.061	0.064	0.065
	BB	0.070	0.079	0.096	0.075	0.070	0.081	0.170	0.151	0.086
	DZ	0.131	0.133	0.130	0.115	0.120	0.116	0.110	0.109	0.109
	PD-DZ ($s=1$)	0.119	0.120	0.118	0.088	0.090	0.088	0.077	0.077	0.073
	PD-DZ ($s \geq 1$)	0.120	0.120	0.119	0.089	0.091	0.089	0.078	0.078	0.075
0.5	AB	0.102	0.135	0.164	0.079	0.089	0.100	0.081	0.084	0.089
	BB	0.086	0.095	0.137	0.102	0.082	0.113	0.231	0.178	0.085
	DZ	0.148	0.135	0.143	0.127	0.131	0.130	0.121	0.121	0.123
	PD-DZ ($s=1$)	0.129	0.117	0.123	0.082	0.083	0.085	0.069	0.067	0.066
	PD-DZ ($s \geq 1$)	0.125	0.116	0.122	0.082	0.081	0.084	0.068	0.070	0.068
0.9	AB	0.440	0.536	0.617	0.249	0.281	0.298	0.144	0.153	0.155
	BB	0.088	0.081	0.092	0.081	0.053	0.082	0.149	0.042	0.068
	DZ	0.151	0.156	0.160	0.137	0.142	0.135	0.131	0.131	0.130
	PD-DZ ($s=1$)	0.108	0.109	0.113	0.075	0.077	0.075	0.058	0.058	0.058
	PD-DZ ($s \geq 1$)	0.097	0.098	0.101	0.064	0.064	0.063	0.045	0.043	0.046

Table 3.4: Monte Carlo simulations: biases and RMSE for all estimators in data with $\varepsilon_{it} \sim$ N$(0,1)$, $\eta_i \sim N(0,1)$, and 10% contamination by independent additive outliers under different sample sizes.

		Bias			RMSE		
	n	100	50	25	100	50	25
α	T	6	12	24	6	12	24
0.1	AB	-0.0761	-0.1063	-0.1360	0.096	0.117	0.142
	BB	-0.0894	-0.1273	-0.2309	0.107	0.139	0.238
	DZ	0.0098	0.0015	0.0072	0.167	0.144	0.137
	PD-DZ ($s=1$)	0.0080	-0.0039	0.0040	0.151	0.114	0.099
	PD-DZ ($s \geq 1$)	0.0083	-0.0026	0.0054	0.152	0.116	0.102
0.5	AB	-0.4728	-0.5090	-0.5321	0.476	0.511	0.534
	BB	-0.4864	-0.5303	-0.6232	0.490	0.533	0.626
	DZ	-0.0173	-0.0183	-0.0148	0.176	0.158	0.151
	PD-DZ ($s=1$)	-0.0183	-0.0186	-0.0124	0.154	0.106	0.086
	PD-DZ ($s \geq 1$)	-0.0186	-0.0167	-0.0071	0.153	0.107	0.088
0.9	AB	-0.8840	-0.9078	-0.9317	0.886	0.909	0.933
	BB	-0.8912	-0.9223	-1.0220	0.893	0.924	1.024
	DZ	-0.0694	-0.0718	-0.0766	0.148	0.141	0.145
	PD-DZ ($s=1$)	-0.0660	-0.0574	-0.0501	0.131	0.096	0.078
	PD-DZ ($s \geq 1$)	-0.0602	-0.0411	-0.0241	0.121	0.079	0.054

Table 3.5: Monte Carlo simulations: biases and RMSE for all estimators in data with $\varepsilon_{it} \sim$ N$(0,1)$, $\eta_i \sim N(0,1)$, and 10% contamination under different sample sizes.

		Bias			RMSE		
	n	100	50	25	100	50	25
α	T	6	12	24	6	12	24
patches of 3 additive outliers							
0.1	AB	0.5713	0.5435	0.5214	0.571	0.545	0.522
	BB	0.5862	0.4861	0.4421	0.594	0.491	0.448
	DZ	0.1965	0.2138	0.2099	0.252	0.255	0.251
	PD-DZ ($s = 1$)	0.2227	0.3272	0.3633	0.267	0.351	0.383
	PD-DZ ($s \geq 1$)	0.2079	0.2910	0.3139	0.248	0.310	0.328
0.5	AB	0.1698	0.1439	0.1215	0.170	0.147	0.125
	BB	0.1877	0.0835	0.0435	0.211	0.107	0.085
	DZ	0.2068	0.2054	0.2115	0.257	0.251	0.249
	PD-DZ ($s = 1$)	0.2165	0.2573	0.2805	0.253	0.275	0.292
	PD-DZ ($s \geq 1$)	0.1294	0.1156	0.1107	0.167	0.135	0.127
0.9	AB	-0.2375	-0.2577	-0.2765	0.240	0.260	0.278
	BB	-0.1858	-0.2975	-0.3493	0.207	0.304	0.356
	DZ	0.0640	0.0686	0.0658	0.096	0.089	0.089
	PD-DZ ($s = 1$)	0.0583	0.0680	0.0689	0.081	0.074	0.072
	PD-DZ ($s \geq 1$)	-0.0420	-0.0363	-0.0222	0.099	0.078	0.057
mix of independent and patches of 3 additive outliers							
0.1	AB	0.2654	0.1987	0.1668	0.302	0.224	0.189
	BB	0.1879	0.1050	-0.0079	0.224	0.144	0.099
	DZ	0.1018	0.1177	0.1031	0.187	0.189	0.172
	PD-DZ ($s = 1$)	0.1157	0.1722	0.1815	0.182	0.211	0.213
	PD-DZ ($s \geq 1$)	0.1125	0.1650	0.1754	0.178	0.201	0.207
0.5	AB	-0.1407	-0.2032	-0.2355	0.206	0.229	0.251
	BB	-0.2196	-0.3005	-0.4089	0.253	0.318	0.422
	DZ	0.1076	0.1114	0.1016	0.196	0.185	0.174
	PD-DZ ($s = 1$)	0.1128	0.1380	0.1457	0.181	0.172	0.169
	PD-DZ ($s \geq 1$)	0.0737	0.0705	0.0726	0.146	0.109	0.101
0.9	AB	-0.5528	-0.6131	-0.6336	0.574	0.622	0.639
	BB	-0.6240	-0.7031	-0.8013	0.637	0.710	0.808
	DZ	0.0101	0.0170	0.0143	0.097	0.085	0.085
	PD-DZ ($s = 1$)	0.0105	0.0262	0.0301	0.082	0.055	0.048
	PD-DZ ($s \geq 1$)	-0.0405	-0.0311	-0.0200	0.099	0.071	0.052

3.4 Monte Carlo simulation

In this section, we evaluate the finite sample performance of the proposed and existing estimators by Monte Carlo simulations. Let $\{y_{it}\}$ follow model (3.1). We generate $T + 100$ observations for each i and discard the first 100 observations to reduce the effect of the initial observations and to achieve stationarity. We consider cases with $\alpha = 0.1, 0.5, 0.9$, $n = 25, 50, 100$, $T = 6, 12, 24$, $\eta_i \sim \mathrm{N}(0, \sigma_\eta^2)$, and $\varepsilon_{it} \sim \mathrm{N}(0, 1)$. If data contamination is present, it follows the contamination schemes (3.12) and (3.13) for $\epsilon = 0.05, 0.10, 0.20$; more specifically, $Z^{1,\epsilon,\zeta}$ uses $G_\zeta = U(10, 90)$ and $Z^{2,\epsilon,\zeta,k}$ employes $p = 3$ and ζ drawn for each patch randomly from $U(10, 90)$; $U(\cdot, \cdot)$ denotes here the uniform distribution. Additionally, we consider a third form of contamination, mixing equally independent additive outliers and patches of outliers. The number of replications is 1000 in all cases. All estimators will be compared by means of the mean bias and the root mean squared error (RMSE).

Results are reported for the original DZ estimator and for the proposed PD-DZ estimator. For the latter, we consider two different specifications of the parameters $\boldsymbol{j} = (s, p, p)' \in \mathcal{J}_o$. Let PD-DZ ($s \geq 1$) denote the PD-DZ estimator when multiple s- and p-differences are used. Note that the theoretical results present above are derived under the assumption that the set of differences is fixed and does not increase with T. Similarly, s and p need to be bounded in simulations. Given the choice of T, it seems reasonable to set $s \in \{1, 3, 5, 7, 9, 11\}$ and $p \in \{1, 3, 5, 7, 9, 11\}$ although a more proper choice of (s, p) needs to be further investigated.

Given the behavior of higher s-differences for small values of the autoregressive parameter, see Figure 3.1, results for the PD-DZ estimator when $s = 1$ and $p \in \{1, 3, 5, 7, 9, 11\}$ are also included. Denote this estimator as PD-DZ ($s = 1$). The standard Arellano-Bond (AB) two-step GMM estimator[2] and the system Blundell and Bond (BB) estimator[3] are also reported, serving as reference estimators. Recall that the AB estimator is strongly negatively biased if α is close to 1, even when data are stationary as is assumed in this paper. The BB estimator, on the contrary, is supposed to perform well under these circumstances.

[2]The (optimal) inverse weight matrix, which is used here, is $\sum_i \boldsymbol{Z}_i^{\mathrm{AB}\prime} \boldsymbol{H} \boldsymbol{Z}_i^{\mathrm{AB}}$, where $\boldsymbol{Z}_i^{\mathrm{AB}}$ is the matrix of instruments and $\boldsymbol{H}$ is a $(T-1) \times (T-1)$ tridiagonal matrix with 2 in the main diagonal, -1 in the first two sub-diagonals, and zeros elsewhere (see Arellano and Bond, 1991, p. 279).

[3]The inverse weight matrix is $\sum_i \boldsymbol{Z}_i^{\mathrm{BB}\prime} \boldsymbol{G} \boldsymbol{Z}_i^{\mathrm{BB}}$, where $\boldsymbol{Z}_i^{\mathrm{BB}}$ is the matrix of instruments and $\boldsymbol{G}$ is a partitioned matrix, $\boldsymbol{G} = \mathrm{diag}(\boldsymbol{H}, \boldsymbol{I})$, where $\boldsymbol{H}$ is as in Arellano-Bond and $\boldsymbol{I}$ is the identity matrix (see Kiviet, 2007, Eq. (38)).

Despite the DZ and PD-DZ estimators being specifically designed to be robust under data contamination, they exhibit a very good performance under clean data with $\sigma_\eta^2 = 1$ as shown in Tables 3.1 and 3.2. Moreover, the proposed PD-DZ estimator exhibits substantially smaller RMSE than the original DZ estimator. It is worth noting that, contrary to AB and BB, the behavior of both DZ and PD-PZ is not very sensitive to the values of the regression parameter. Finally, the performance of PD-DZ is rather close to that of the BB estimator, which is the estimator with the overall best performance; the differences are small especially for large T and large α. Note that the unreported simulations indicate that the results do differ only slightly if the error distribution follows the Laplace or Student distributions. A similar picture results from Table 3.3, where the performance of the estimators for different values of the ratio $\sigma_\eta^2/\sigma_\varepsilon^2$ is reported (i.e., for $\sigma_\eta^2 = 1/4$, 1, and 4 and $\sigma_\varepsilon^2 = 1$). Contrary to BB, the behavior of both DZ and PD-DZ does not significantly change with the ratio $\sigma_\eta^2/\sigma_\varepsilon^2$. The proposed PD-DZ is now always preferable to BB except for the case of small $\alpha = 0.1$ and small $T \leq 12$.

Next, three different data contaminations schemes are considered: independent additive outliers in Table 3.4, patches of additive outliers and a mix of independent and patches of outliers in Table 3.5. In the range of $\epsilon = 0.05, \ldots, 0.20$, the bias and RMSE results are approximately proportional to ϵ and we thus report only the case with $\epsilon = 0.10$. As expected, AB and BB are strongly biased in all cases. In the case of independent additive outliers, the negative biases of DZ and PD-DZ are rather small and the proposed PD-DZ estimator performs better than DZ both in terms of the bias and RMSE. On the other hand, the patches of additive outliers lead to positive biases, which are rather small for $\alpha = 0.9$, but which increase quickly as α decreases. Although PD-DZ is preferable to DZ for $\alpha \geq 0.5$, PD-DZ is much more influenced by the patches of outliers if the autoregressive parameter α is close to zero. The same conclusion holds for the case of a mixed form of contamination.

3.5 Concluding remarks

In this paper, we propose an extension of the median-based robust estimator for dynamic panel data model of Dhaene and Zhu (2009) by means of multiple pairwise differences. This yields an estimator which is more efficient and more globally robust. The main concern is the larger bias of PD-DZ in the presence of patches of additive outliers if α is close to zero. This could possibly be solved by means of weights which are inversely proportional to the bias and will be a topic of further research.

3.A Appendix

Let us first derive formula (3.7) for r_j given its definition in (3.3). The covariance between $\Delta^s y_{it}$ and $\Delta^p y_{it-q}$ can be decomposed as

$$\begin{aligned}\operatorname{cov}(\Delta^s y_{it}, \Delta^p y_{it-q}) &= \operatorname{cov}(y_{it}, y_{it-q}) - \operatorname{cov}(y_{it}, y_{it-q-p}) \\ &\quad - \operatorname{cov}(y_{it-s}, y_{it-q}) + \operatorname{cov}(y_{it-s}, y_{it-q-p}).\end{aligned} \tag{3.31}$$

Next, a general expression for $\operatorname{cov}(y_{it}, y_{it-q})$ has to be derived, which is time independent due to the stationarity of y_{it}. Given Assumptions A.1 and A.2, let σ_η^2 and σ_y^2 denote the variance of η_i and y_{it}, respectively, and let $\sigma_{y\eta} = \operatorname{cov}(y_{it}, \eta_i)$ (this covariance is again time independent because of Assumption A.2). Thus each of the terms in (3.31) can be generically expressed as (using recursive substitution for y_{it} from the model equation (3.1))

$$\begin{aligned}\operatorname{cov}(y_{it}, y_{it-q}) &= \operatorname{cov}\left\{\left[\left(\sum_{k=0}^{q-1}\alpha^k\right)\eta_i + \alpha^q y_{it-q} + \sum_{k=0}^{q-1}\alpha^k \varepsilon_{it-k}\right], y_{it-q}\right\} \\ &= \left(\sum_{k=0}^{q-1}\alpha^k\right)\sigma_{y\eta} + \alpha^q\sigma_y^2 = \frac{\sigma_{y\eta}^2}{1-\alpha} + \left(\sigma_y^2 - \frac{\sigma_{y\eta}}{1-\alpha}\right)\alpha^q.\end{aligned} \tag{3.32}$$

It follows that $\operatorname{cov}(\Delta^s y_{it}, \Delta^p y_{it-q}) = c\left(\alpha^q - \alpha^{q+p} - \alpha^{|s-q|} + \alpha^{|s-p-q|}\right)$, where $c := \sigma_y^2 - \sigma_{y\eta}/(1-\alpha)$. The result in (3.7) now follows directly by the application of the same steps to $\operatorname{var}(\Delta^p y_{it-q})$:

$$\begin{aligned}\operatorname{var}(\Delta^p y_{it-q}) &= \operatorname{cov}(y_{it-q}, y_{it-q}) - \operatorname{cov}(y_{it-q}, y_{it-q-p}) \\ &\quad - \operatorname{cov}(y_{it-q-p}, y_{it-q}) + \operatorname{cov}(y_{it-q-p}, y_{it-q-p}) \\ &= c(1 - \alpha^p - \alpha^p + 1) = 2c(1-\alpha^p).\end{aligned} \tag{3.33}$$

3.A.1 Asymptotic distribution

Lemma 2. *Let the moment conditions in* (3.9) *be considered for a finite set $\mathcal{J}$ of indices and Assumptions A.1–A.4 hold. Then for a fixed T and $n \to \infty$*

$$\sqrt{n}\boldsymbol{g}_n(\alpha) \to \mathrm{N}(\boldsymbol{0}, \boldsymbol{V}), \tag{3.34}$$

where the length $l(\boldsymbol{j}) = \max\{s, p+q\}$ and the square matrix $\boldsymbol{V}$ has a typical element with indices $\boldsymbol{j}, \boldsymbol{j}' \in \mathcal{J}$

$$\begin{aligned} v_{\boldsymbol{j}\boldsymbol{j}'} = \pi^2 \sqrt{e_{\boldsymbol{j}\boldsymbol{j}'}}\, \mathrm{E} \Bigg[\Bigg(\sum_{t=l(\boldsymbol{j})+1}^{T} \operatorname{sgn}(\Delta^s y_{it} - r_{\boldsymbol{j}} \Delta^p y_{it-q}) \operatorname{sgn}(\Delta^p y_{it-q}) \Bigg) \times \\ \Bigg(\sum_{t=l(\boldsymbol{j}')+1}^{T} \operatorname{sgn}(\Delta^{s'} y_{it} - r_{\boldsymbol{j}'} \Delta^{p'} y_{it-q'}) \operatorname{sgn}(\Delta^{p'} y_{it-q'}) \Bigg) \Bigg], \end{aligned} \tag{3.35}$$

and

$$\begin{aligned} e_{\boldsymbol{j}\boldsymbol{j}'} = & \left(1 - \alpha^s - \frac{1}{4}[\alpha^q - \alpha^{p+q} - \alpha^{|s-q|} + \alpha^{|s-p-q|}]^2\right) \times \\ & \left(1 - \alpha^{s'} - \frac{1}{4}[\alpha^{q'} - \alpha^{p'+q'} - \alpha^{|s'-q'|} + \alpha^{|s'-p'-q'|}]^2\right) \times \\ & \{[T - l(\boldsymbol{j})][T - l(\boldsymbol{j}')]\}^{-1} \end{aligned} \tag{3.36}$$

Proof. To prove the lemma, note that $\boldsymbol{g}_{n\boldsymbol{j}}(\alpha)$ can be rewritten in the following way for any index $\boldsymbol{j}$:

$$\begin{aligned} \sqrt{n} g_{n\boldsymbol{j}}(\alpha) &= \sqrt{n} \left[2(1-\alpha^p) r_{\boldsymbol{j}} - \alpha^q + \alpha^{q+p} + \alpha^{|s-q|} - \alpha^{|s-p-q|}\right] \\ &= \sqrt{n} 2(1-\alpha^p) \left(\hat{r}_{n\boldsymbol{j}} - \frac{\alpha^q - \alpha^{q+p} - \alpha^{|s-q|} + \alpha^{|s-p-q|}}{2(1-\alpha^p)} \right) \\ &= \sqrt{n} 2(1-\alpha^p) \left(\hat{r}_{n\boldsymbol{j}} - r_{\boldsymbol{j}}\right) \\ &= \frac{2(1-\alpha^p)}{\sqrt{T}} \sqrt{nT} (\hat{r}_{n\boldsymbol{j}} - r_{\boldsymbol{j}}). \end{aligned} \tag{3.37}$$

Since the proof of Dhaene and Zhu (2009, Lemma 1) is valid not only for the first differences, but for any sth difference, we can use the results derived in the proof of Lemma 1 in Dhaene and Zhu (2009) to state that

$(\sqrt{nT}[\hat{r}_{\boldsymbol{j}} - r_{\boldsymbol{j}}])_{\boldsymbol{j}\in\mathcal{J}}$ has the same asymptotic distribution as

$$\left(\frac{\sqrt{T}}{\sqrt{T - l(\boldsymbol{j})}} \frac{a_{\boldsymbol{j}}(r_{\boldsymbol{j}})}{\sqrt{n[T - l(\boldsymbol{j})]}} \sum_{i=1}^{n} \sum_{t=l(\boldsymbol{j})+1}^{T} \operatorname{sgn}(\Delta^s y_{it} - r_{\boldsymbol{j}} \Delta^p y_{it-q}) \operatorname{sgn}(\Delta^p y_{it-q}) \right)_{\boldsymbol{j}\in\mathcal{J}},$$

where the constant $a_{\boldsymbol{j}}(r_{\boldsymbol{j}}) = f_{i\boldsymbol{j}}(0) \operatorname{E} |\Delta^s y_{it-p}|/2 = \pi/2\sqrt{\operatorname{var}(\Delta^s y_{it} - r_{\boldsymbol{j}}\Delta^p y_{it-q})/\operatorname{var}(\Delta^p y_{it-q})}$ and $f_{i\boldsymbol{j}}$ denotes the density function of $\Delta^s y_{it} - r_{\boldsymbol{j}}\Delta^p y_{it-q}$.

To derive the expression for $a_{\boldsymbol{j}}(r_{\boldsymbol{j}})$, note that the variables $\Delta^s y_{it} - r_{\boldsymbol{j}}\Delta^p y_{it-s}$ and $\Delta^p y_{it-s}$ are uncorrelated because of definition (3.2), and by Assumption A.3, they are also independent and normally distributed around zero. From Equation (3.33), it follows that

$$\begin{pmatrix} \Delta^s y_{it} - r_{\boldsymbol{j}}\Delta^p y_{it-q} \\ \Delta^p y_{it-q} \end{pmatrix} \sim \operatorname{N}\left[\mathbf{0}, 2c \begin{pmatrix} 1 - \alpha^s - r_{\boldsymbol{j}}^2(1-\alpha^p) & 0 \\ 0 & 1-\alpha^p \end{pmatrix}\right] \tag{3.38}$$

because equation (3.3) implies $\operatorname{cov}(\Delta^s y_{it}, \Delta^p y_{it-q}) = r_{\boldsymbol{j}} \operatorname{var}(\Delta^p y_{it-q})$ and thus

$$\begin{aligned} \operatorname{var}(\Delta^s y_{it} - r_{\boldsymbol{j}}\Delta^p y_{it-q}) &= \operatorname{var}(\Delta^s y_{it}) + r_{\boldsymbol{j}}^2 \operatorname{var}(\Delta^p y_{it-q}) - 2r_{\boldsymbol{j}} \operatorname{cov}(\Delta^s y_{it}, \Delta^p y_{it-q}) \\ &= \operatorname{var}(\Delta^s y_{it}) - r_{\boldsymbol{j}}^2 \operatorname{var}(\Delta^p y_{it-s}). \end{aligned} \tag{3.39}$$

Hence, substituting from (3.7) for $r_{\boldsymbol{j}}$ implies

$$a_{\boldsymbol{j}}(r_{\boldsymbol{j}}) = \frac{\pi}{2}\sqrt{(1 - \alpha^s - [\alpha^q - \alpha^{p+q} - \alpha^{|s-q|} + \alpha^{|s-p-q|}]^2/4)/(1-\alpha^p)^2}.$$

The claim of the lemma now follows from the application of the central limit theorem to (3.38) with respect to $n \to \infty$ because the denominator of $a_{\boldsymbol{j}}(r_{\boldsymbol{j}})$ cancels with the coefficient $2(1-\alpha^p)$ in (3.37).

□

Proof of Theorem 3. The estimator $\hat{\alpha}_n$ is defined by the solution of the sample analogs of equations (3.9), which are deterministic functions of $\hat{r}_{n\boldsymbol{j}}$. Thus the stochastic behavior of the moment equations is fully determined by the asymptotic properties of $\hat{r}_{n\boldsymbol{j}}$, which are given in Lemma 2. To derive the asymptotic distribution of $\hat{\alpha}_n$, we will thus use general consistency and asymptotic normality theorems given, for example, in Arellano (2003, Appendix A.4 and A.5). Given that by assumption the true parameter $\alpha \in (-1,1)$, moment conditions $\boldsymbol{g}_n(\alpha)$ are differentiable with respect to α. Denote $\boldsymbol{d}_n(\alpha) := \partial\boldsymbol{g}_n(\alpha)/\partial\alpha$. The GMM estimator $\hat{\alpha}_n$ satisfies the first-order conditions

$$\boldsymbol{d}_n'(\hat{\alpha}_n)\boldsymbol{A}_n\boldsymbol{g}_n(\hat{\alpha}_n) = 0. \tag{3.40}$$

The derivatives $\boldsymbol{d}_n(\alpha)$ converge to $\boldsymbol{d} := \boldsymbol{d}(\alpha) = \partial\boldsymbol{g}(\alpha)/\partial\alpha$ uniformly in α because their only stochastic element $\hat{r}_{n\boldsymbol{j}}$ is independent of α. Together with Assumption (A.4), it follows that

$$\boldsymbol{d}'\boldsymbol{A}\sqrt{n}\boldsymbol{g}_n(\hat{\alpha}_n) = o_p(1). \tag{3.41}$$

Using the first-order expansion, we have

$$\boldsymbol{d}'\boldsymbol{A}\left[\sqrt{n}\boldsymbol{g}_n(\alpha) + \boldsymbol{d}_n(\alpha)(\hat{\alpha}_n - \alpha)\right] = o_p(1), \tag{3.42}$$

that is,

$$\sqrt{n}(\hat{\alpha}_n - \alpha) = -(\boldsymbol{d}'\boldsymbol{A}\boldsymbol{d})^{-1}\boldsymbol{d}'\boldsymbol{A}\sqrt{n}\boldsymbol{g}_n(\alpha) + o_p(1). \tag{3.43}$$

Because the moment conditions $\boldsymbol{g}(\alpha) = 0$ are continuous functions of α, $\hat{r}_{n\boldsymbol{j}}$ is converges to $r_{\boldsymbol{j}}$ (uniformly in α as it is independent of α and uniformly in $\boldsymbol{j}$ as $\mathcal{J}$ is finite), and Assumption A.4 holds, the consistency of $\hat{\alpha}_n$ follows from the theorem in Arellano (2003, Appendix A.4) if the true parameter α is uniquely identified. This however directly follows from $(1,1,1)' \in \mathcal{J}$ because this first equation corresponding to $g_{111}(\alpha) = (1-\alpha)(2r_{111} + 1 - \alpha) = 0$ has only one solution $\alpha = 1 + 2r_{111}$ on $(-1, 1)$.

Moreover, Lemma 2 guarantees the asymptotic normality of the moment conditions $\boldsymbol{g}_n(\alpha)$. We can thus use the asymptotic normality theorem in Arellano (2003, Appendix A.5, pp. 187), which states that

$$\sqrt{n}(\hat{\alpha}_n - \alpha) \to \mathrm{N}(0, (\boldsymbol{d}'\boldsymbol{A}\boldsymbol{d})^{-1}\boldsymbol{d}'\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{d}(\boldsymbol{d}'\boldsymbol{A}\boldsymbol{d})^{-1}) \tag{3.44}$$

as $n \to \infty$, where $\boldsymbol{V}$ is derived in Lemma 2. □

3.A.2 Robustness properties

The breakdown properties stated in Theorems 4–6 are derived similarly to Theorems 5 and 8 stated in Dhaene and Zhu (2009). In order to prove the theorems concerning the influence function and the breakdown point of $\hat{\alpha}$, it is useful to derive first the asymptotic bias of $\hat{r}_{\boldsymbol{j}}$ as an estimator of $r_{\boldsymbol{j}}$.

Similarly to (3.23), this is defined as

$$\operatorname{bias}\left(\hat{r}_{\boldsymbol{j}}; Z^{\alpha}, Z^{\mathfrak{I},\epsilon,\zeta}\right) := \operatorname{plim}_{n\to\infty} \hat{r}_{\boldsymbol{j}}(Z^{\alpha}_{nT} + Z^{\mathfrak{I},\epsilon,\zeta}_{nT}) - r_{\boldsymbol{j}}, \tag{3.45}$$

where plim denotes the probability limit operator. Let $b := b(r_{\boldsymbol{j}}, \zeta, \epsilon_1)$ be a short-hand notation for (3.45). Then, b solves the following equation

$$\mathrm{E}\left[\operatorname{sgn}\left(\frac{\Delta^s y^{\epsilon,\zeta}_{it}}{\Delta^p y^{\epsilon,\zeta}_{it-s}} - r_{\boldsymbol{j}}\right)\right] = b, \tag{3.46}$$

which can be written also as

$$\Pr\left(\frac{\Delta^s y^{\epsilon,\zeta}_{it} - r_{\boldsymbol{j}}\Delta^p y^{\epsilon,\zeta}_{it-s}}{\Delta^p y^{\epsilon,\zeta}_{it-s}} \leq b\right) = \frac{1}{2}, \tag{3.47}$$

where $y^{\epsilon,\zeta}_{it} = y_{it} + z^{\epsilon}_{it}$, z^{ϵ}_{it} follows one of the three data contamination processes in (3.12)–(3.14). Since $r_{\boldsymbol{j}}$ is considered only for $\boldsymbol{j} = (s, s, p)' \in \mathcal{J}_o$, where both s and p are odd, $r_{\boldsymbol{j}} = -(1-\alpha^s)/2$. This mapping of α to $r_{\boldsymbol{j}} = -(1-\alpha^s)/2$ has the same important properties for $s = 1$ and any odd $s > 1$: it maps interval $(-1, 0)$ to $(-1, -1/2)$ and interval $(0, 1)$ to $(-1/2, 0)$, it is continuous, and it is strictly increasing on $(-1, 1)$. One can thus follow the proofs in Dhaene and Zhu (2009, Theorems 5 and 8) and apply them not only to the case of $s = p = 1$, but any odd s and p with only two adjustments: (i) the variables $\Delta^s y_{it} - r_{\boldsymbol{j}}\Delta^p y_{it-s}$ and $\Delta^p y_{it-s}$ have to be standardized (Dhaene and Zhu, 2009, equation (17)) and their variances generally depend on the values of s and p, and (ii) in the case of patches of outliers, the probability that a patch contaminates the ratio $\Delta^s y_{it}/\Delta^p y_{it-s}$ needs to be generalized.

As for (i), note that by Equation (3.2) the variables $\Delta^s y_{it} - r_{\boldsymbol{j}}\Delta^p y_{it-s}$ and $\Delta^p y_{it-s}$ are uncorrelated, and by Assumption A.3, they are independent and normally distributed around zero. From Equation (3.33), it follows that

$$\begin{pmatrix}\Delta^s y_{it} - r_{\boldsymbol{j}}\Delta^p y_{it-s} \\ \Delta^p y_{it-s}\end{pmatrix} \sim \mathrm{N}\left[\mathbf{0}, \frac{2\sigma^2_{\varepsilon}}{1-\alpha^2}\begin{pmatrix}1-\alpha^s - r^2_{\boldsymbol{j}}(1-\alpha^p) & 0 \\ 0 & 1-\alpha^p\end{pmatrix}\right] \tag{3.48}$$

because Equation (3.3) implies $\operatorname{cov}(\Delta^s y_{it}, \Delta^p y_{it-s}) = r_{\boldsymbol{j}} \operatorname{var}(\Delta^p y_{it-s})$ and thus

$$\begin{aligned}\operatorname{var}(\Delta^s y_{it} - r_{\boldsymbol{j}} \Delta^p y_{it-s}) &= \operatorname{var}(\Delta^s y_{it}) + r_{\boldsymbol{j}}^2 \operatorname{var}(\Delta^p y_{it-s}) - 2r_{\boldsymbol{j}} \operatorname{cov}(\Delta^s y_{it}, \Delta^p y_{it-s}) \\ &= \operatorname{var}(\Delta^s y_{it}) - r_{\boldsymbol{j}}^2 \operatorname{var}(\Delta^p y_{it-s}).\end{aligned} \tag{3.49}$$

Independent additive outlier contamination $Z^{1,\epsilon,\zeta}$

Under independent additive outlier contamination $Z^{1,\epsilon,\zeta}$, Equation (3.47) can be written as

$$\begin{aligned}\Pr\left(\frac{\Delta^s y_{it}^{\epsilon,\zeta} - r_{\boldsymbol{j}} \Delta^p y_{it-s}^{\epsilon,\zeta}}{\Delta^p y_{it-s}^{\epsilon,\zeta}} \leq b\right) &= \Pr\left(\frac{u_{it\boldsymbol{j}} + \Delta^s z_{it}^{\epsilon} - r_{\boldsymbol{j}} \Delta^p z_{it-s}^{\epsilon}}{\Delta^p y_{it-s} + \Delta^p z_{it-s}^{\epsilon}} \leq b\right) \\ &= \Pr\left[f(\boldsymbol{z}_{it}^{\epsilon}) \leq b\right] = \frac{1}{2},\end{aligned} \tag{3.50}$$

where $\boldsymbol{z}_{it}^{\epsilon} = (z_{it}^{\epsilon}, z_{it-s}^{\epsilon}, z_{it-s-p}^{\epsilon})'$ is a random vector following (3.12), hence $f(\boldsymbol{z}_{it}^{\epsilon})$ is a random scalar. Let $\Omega_{\boldsymbol{z}_{it}^{\epsilon}}$ be the set of the eight possible outcomes of $\boldsymbol{z}_{it}^{\epsilon}$, that is

$$\Omega_{\boldsymbol{z}_{it}^{\epsilon}} := \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ \zeta \end{pmatrix}, \cdots, \begin{pmatrix} \zeta \\ \zeta \\ \zeta \end{pmatrix} \right\}, \tag{3.51}$$

where the number of elements is $\#\Omega_{\boldsymbol{z}_{it}^{\epsilon}} = 8$. To simplify the notation, let us refer to (3.51) as Ω_{it}, and denote each of its element as $\boldsymbol{\omega}_{itj}$, $j = 1, \cdots, 8$. Then it holds

$$\begin{aligned}\Pr\left[f(\boldsymbol{z}_{it}^{\epsilon}) \leq b\right] &= \Pr\left[f(\boldsymbol{z}_{it}^{\epsilon}) \leq b \cap \left(\bigcup_{j=1}^{8} \boldsymbol{z}_{it}^{\epsilon} = \boldsymbol{\omega}_{itj}\right)\right] \\ &= \sum_{j=1}^{8} \Pr\left[f(\boldsymbol{z}_{it}^{\epsilon}) \leq b \cap (\boldsymbol{z}_{it}^{\epsilon} = \boldsymbol{\omega}_{itj})\right] \\ &= \sum_{j=1}^{8} \Pr\left[f(\boldsymbol{z}_{it}^{\epsilon}) \leq b \,|\boldsymbol{z}_{it}^{\epsilon} = \boldsymbol{\omega}_{itj}\right] \Pr\left(\boldsymbol{z}_{it}^{\epsilon} = \boldsymbol{\omega}_{itj}\right) \\ &= \sum_{j=1}^{8} \Pr\left[f(\boldsymbol{\omega}_{itj}) \leq b\right] \Pr\left(\boldsymbol{z}_{it}^{\epsilon} = \boldsymbol{\omega}_{itj}\right).\end{aligned} \tag{3.52}$$

Note that $\Pr\left(\boldsymbol{z}_{it}^{\epsilon} = \boldsymbol{\omega}_{itj}\right) = \Pr\left(\boldsymbol{z}_{it}^{\epsilon} = \boldsymbol{\omega}_{itj'}\right)$ for some j and j', because the data contamination $Z^{1,\epsilon,\zeta}$ is characterized by outliers occurring independently from each other. For instance, $\Pr[(\zeta, 0, 0)'] = \Pr[(0, \zeta, 0)'] = \Pr[(0, 0, \zeta)'] =$

$(1-\epsilon_1)^2\epsilon_1$. Moreover, $f[(0,0,0)'] = f[(\zeta,\zeta,\zeta)']$. Therefore, Equation (3.50) can be decomposed as

$$\Pr\left[f(\boldsymbol{z}_{it}^{\epsilon}) \leq b\right] = \left[(1-\epsilon_1)^3 + \epsilon_1^3\right] A + (1-\epsilon_1)^2\epsilon_1 B + (1-\epsilon_1)\epsilon_1^2 C = \frac{1}{2}, \tag{3.53}$$

where

$$\begin{aligned} A(r_{\boldsymbol{j}}, b) &:= \Pr\left(\frac{u_{it\boldsymbol{j}}}{\Delta^p y_{it-s}} \leq b\right) \\ B(r_{\boldsymbol{j}}, \zeta, b) &:= \Pr\left(\frac{u_{it\boldsymbol{j}} + \zeta}{\Delta^p y_{it-s}} \leq b\right) + \Pr\left(\frac{u_{it\boldsymbol{j}} - \zeta(1+r_{\boldsymbol{j}})}{\Delta^p y_{it-s} + \zeta} \leq b\right) + \Pr\left(\frac{u_{it\boldsymbol{j}} + \zeta r_{\boldsymbol{j}}}{\Delta^p y_{it-s} - \zeta} \leq b\right) \\ C(r_{\boldsymbol{j}}, \zeta, b) &:= \Pr\left(\frac{u_{it\boldsymbol{j}} - \zeta r_{\boldsymbol{j}}}{\Delta^p y_{it-s} + \zeta} \leq b\right) + \Pr\left(\frac{u_{it\boldsymbol{j}} + \zeta(1+r_{\boldsymbol{j}})}{\Delta^p y_{it-s} - \zeta} \leq b\right) + \Pr\left(\frac{u_{it\boldsymbol{j}} - \zeta}{\Delta^p y_{it-s}} \leq b\right). \end{aligned} \tag{3.54}$$

These probabilities are of the form

$$L(k, l, b) = \Pr\left(\frac{u_{it\boldsymbol{j}} + k}{\Delta^p y_{it-s} - l} \leq b\right), \tag{3.55}$$

for given k, l, and b and they can be conveniently standardised by using (3.48) as follows

$$L(k, l, b) = \Pr\left(\frac{X + k'}{Y - l'} \leq b'\right), \tag{3.56}$$

where X and Y are independent $\mathrm{N}(0,1)$ variables and

$$k' := \frac{k}{\sigma_u}, \qquad l' := \frac{l}{\sigma_{\Delta^p}}, \qquad b' := \sigma^* b \tag{3.57}$$

and

$$\sigma^* := \frac{\sigma_{\Delta^p}}{\sigma_u} = \sqrt{\frac{1-\alpha^p}{1-\alpha^s-(1-\alpha^s)^2(1-\alpha^p)/4}}, \tag{3.58}$$

where $\sigma_u := \sqrt{\mathrm{var}(u_{it\boldsymbol{j}})}$ and $\sigma_{\Delta^p} := \sqrt{\mathrm{var}(\Delta^p y_{it-s})}$ as in (3.48). Finally, note that $L(k,l,b) = L(-k,-l,b)$, hence $B = C$ and (3.50) becomes

$$A + \epsilon_1(1-\epsilon_1)(B - 3A) = \frac{1}{2}. \tag{3.59}$$

Proof of Theorem 4. This result is direct consequence of Theorem 5 stated in Dhaene and Zhu (2009). In this case, the standardization of b' in Dhaene and Zhu (2009, Equation (17)) is done as in (3.58) by using the result in (3.38) , and correspond to the factor

$$\frac{1-\alpha^s}{1-\alpha^p} - r_{\boldsymbol{j}}^2 = \frac{1-\alpha^s}{1-\alpha^p} - \frac{1}{4}(1-\alpha^s)^2, \tag{3.60}$$

instead of to factor $1 - r_{111}^2$ in the original paper. This concludes the proof once we substitute $-(1-\alpha^s)/2$ for $r_{\boldsymbol{j}}$ in Dhaene and Zhu (2009, Theorem 5). □

Proof of Theorem 7. It follows from the definition of influence function that

$$\mathrm{IF}(\hat{r}_{\boldsymbol{j}}; Z^\alpha, Z^{1,\epsilon,\zeta}) := \left.\frac{\partial\,\mathrm{bias}(\hat{r}_{\boldsymbol{j}}; Z^\alpha, Z^{1,\epsilon,\zeta})}{\partial\epsilon_1}\right|_{\epsilon_1=0} = \frac{3A(r_{\boldsymbol{j}},0) - B(r_{\boldsymbol{j}},\zeta,0)}{A_b'(r_{\boldsymbol{j}},0)}, \tag{3.61}$$

where the equality follows from the implicit function theorem applied to (3.59) and where

$$A_b'(r_{\boldsymbol{j}},0) := \left.\frac{\partial A(r_{\boldsymbol{j}},b)}{\partial b}\right|_{\epsilon=0}. \tag{3.62}$$

Therefore, similarly to Dhaene and Zhu (2009, Equation (18)),

$$A(r_{\boldsymbol{j}},b) = \Pr\left(\frac{X}{Y} \le \sigma^* b\right) = \frac{1}{2} + \frac{1}{\pi}\arctan\sigma^* b, \tag{3.63}$$

where σ^* is defined in (3.58). Hence, $A(r_{\boldsymbol{j}},0) = 1/2$ and

$$A_b'(r_{\boldsymbol{j}},0) = \frac{1}{\pi}\sqrt{\frac{1-\alpha^p}{1-\alpha^s - r_{\boldsymbol{j}}^2(1-\alpha^p)}}. \tag{3.64}$$

Next,

$$\begin{aligned} B(r_{\boldsymbol{j}},\zeta,0) = \frac{3}{2} + &\left[\Phi\left(\frac{\zeta(1+r_{\boldsymbol{j}})}{\sqrt{2c[1-\alpha^s - r_{\boldsymbol{j}}^2(1-\alpha^p)]}}\right) - \Phi\left(-\frac{\zeta r_{\boldsymbol{j}}}{\sqrt{2c[1-\alpha^s - r_{\boldsymbol{j}}^2(1-\alpha^p)]}}\right)\right] \\ &\times\left[\Phi\left(\frac{\zeta}{\sqrt{2c(1-\alpha^p)}}\right) - \Phi\left(-\frac{\zeta}{\sqrt{2c(1-\alpha^p)}}\right)\right]. \end{aligned} \tag{3.65}$$

by Dhaene and Zhu (2009, Lemma 3). Replacing $c = \sigma_\varepsilon^2/(1-\alpha^2)$ and

$r_{\boldsymbol{j}} = -(1-\alpha^s)/2$ completes the proof. □

Patch additive outlier contamination $Z^{2,\epsilon,\zeta,k}$

As in Section 3.A.2, it is useful to derive first the asymptotic bias of $\hat{r}_{\boldsymbol{j}}$ under the outlier contamination $Z^{2,\epsilon,\zeta,k}$ as defined in (3.13). Similarly to (3.45), this is given by $b := b(r_{\boldsymbol{j}}, \zeta, \epsilon_1, k)$ solving the equation

$$\begin{aligned}\Pr\left(\frac{\Delta^s y_{it}^{\epsilon,\zeta} - r_{\boldsymbol{j}}\Delta^p y_{it-s}^{\epsilon,\zeta}}{\Delta^p y_{it-s}^{\epsilon,\zeta}} \le b\right) &= \Pr\left(\frac{u_{it\boldsymbol{j}} + \Delta^s z_{it}^{\epsilon} - r_{\boldsymbol{j}}\Delta^p z_{it-s}^{\epsilon}}{\Delta^p y_{it-s} + \Delta^p z_{it-s}^{\epsilon}} \le b\right) \\ &= \mathfrak{p}_A A + \mathfrak{p}_B B + \mathfrak{p}_C C + \mathfrak{p}_D D = \frac{1}{2},\end{aligned} \tag{3.66}$$

where the notation is defined below. Note that the second equality follows along the same line as in Section 3.A.2, in particular Equation (3.52). In this case, the only difference is that outliers no longer occur independently but in patches. The number of elements of Ω_{it} increases to $\#\Omega_{it} = 13$ as now, if we observe multiple outliers, we shall distinguish the event of them belonging to the same patch from the event of these outliers belonging to different patches. For instance, $(0,\zeta,\zeta)'$ may be that result of one patch only, $(0,\zeta_1,\zeta_1)'$, or of two patches, $(0,\zeta_2,\zeta_1)'$. Hence,

$$\begin{aligned}\mathfrak{p}_B &:= \Pr\left[\begin{pmatrix}\zeta\\0\\0\end{pmatrix} \cup \begin{pmatrix}0\\\zeta\\\zeta\end{pmatrix}\right] = \Pr\begin{pmatrix}\zeta_1\\0\\0\end{pmatrix} + \Pr\begin{pmatrix}0\\\zeta_1\\\zeta_1\end{pmatrix} + \Pr\begin{pmatrix}0\\\zeta_2\\\zeta_1\end{pmatrix} \\ &= (1-\epsilon_2)^{k+\min\{p,k\}} \cdot \epsilon_2 \cdot \min\{s,k\} \\ &\quad + \epsilon_2 \cdot \max\Big\{0, s+k-\max\{s+p,k\}\Big\} \cdot (1-\epsilon_2)^k \\ &\qquad + \epsilon_2^2 \cdot \Big(p+k-\max\{p,k\}\Big) \cdot \max\Big\{0, s+\min\{p,k\}-\max\{s,k\}\Big\} \cdot (1-\epsilon_2)^k,\end{aligned} \tag{3.67}$$

$$\begin{aligned}\mathfrak{p}_C &:= \Pr\left[\begin{pmatrix}0\\0\\\zeta\end{pmatrix} \cup \begin{pmatrix}\zeta\\\zeta\\0\end{pmatrix}\right] = \Pr\begin{pmatrix}0\\0\\\zeta_1\end{pmatrix} + \Pr\begin{pmatrix}\zeta_1\\\zeta_1\\0\end{pmatrix} + \Pr\begin{pmatrix}\zeta_2\\\zeta_1\\0\end{pmatrix} \\ &= \epsilon_2 \cdot \Big(p+k-\max\{p,k\}\Big) \cdot (1-\epsilon_2)^{k+\min\{s,k\}} \\ &\quad + (1-\epsilon_2)^k \cdot \epsilon_2 \cdot \max\Big\{0, \min\{s+p,k\}-s\Big\} \\ &\qquad + (1-\epsilon_2)^k \cdot \epsilon_2^2 \cdot \max\Big\{0, s+\min\{p,k\}-\max\{s,k\}\Big\} \cdot \min\{s,k\},\end{aligned}$$

$$(3.68)$$

$$\begin{aligned}\mathfrak{p}_D := \Pr\left[\begin{pmatrix}\zeta\\0\\\zeta\end{pmatrix}\cup\begin{pmatrix}0\\\zeta\\0\end{pmatrix}\right] &= \Pr\begin{pmatrix}\zeta_2\\0\\\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}0\\\zeta_1\\0\end{pmatrix}\\ &= \epsilon_2^2\cdot\Big(p+k-\max\{p,k\}\Big)\cdot(1-\epsilon_2)^k\cdot\min\{s,k\}\\ &\quad+(1-\epsilon_2)^{2k}\cdot\epsilon_2\cdot\max\Big\{0,s+\min\{p,k\}-\max\{s,k\}\Big\},\end{aligned} \tag{3.69}$$

and $\mathfrak{p}_A = 1-\mathfrak{p}_B-\mathfrak{p}_C-\mathfrak{p}_D$. Next,

$$\begin{aligned}A(r_{\boldsymbol{j}},b) &:= \Pr\left(\frac{u_{it\boldsymbol{j}}}{\Delta^p y_{it-s}}\le b\right)\\ B(r_{\boldsymbol{j}},\zeta,b) &:= \Pr\left(\frac{u_{it\boldsymbol{j}}+\zeta}{\Delta^p y_{it-s}}\le b\right)\\ C(r_{\boldsymbol{j}},\zeta,b) &:= \Pr\left(\frac{u_{it\boldsymbol{j}}+\zeta r_{\boldsymbol{j}}}{\Delta^p y_{it-s}-\zeta}\le b\right)\\ D(r_{\boldsymbol{j}},\zeta,b) &:= \Pr\left(\frac{u_{it\boldsymbol{j}}+\zeta(1+r_{\boldsymbol{j}})}{\Delta^p y_{it-s}-\zeta}\le b\right)\end{aligned} \tag{3.70}$$

where the symmetry $L(k,l,b) = L(-k,-l,b)$ has been used, recall Equation (3.55).

Proof of Theorem 5. Denote the left hand side of (3.66) as $V(r_{\boldsymbol{j}},\zeta,b,k,\epsilon_2)$. As $\zeta\to\infty$, $\hat{r}_{\boldsymbol{j}}$ breaks down to 0 if and only if

$$\lim_{b\uparrow -r_{\boldsymbol{j}}}\lim_{\zeta\to\infty} V(r_{\boldsymbol{j}},\zeta,b,k,\epsilon_2)\le\frac{1}{2}. \tag{3.71}$$

Next

$$\begin{aligned}\lim_{b\uparrow -r_{\boldsymbol{j}}}\lim_{\zeta\to\infty} B(r_{\boldsymbol{j}},\zeta,b) &= \lim_{b\uparrow -r_{\boldsymbol{j}}}\lim_{\zeta\to\infty}\Pr\left(\frac{\Delta^s y_{it}+\zeta}{\Delta^p y_{it-s}}\le r_{\boldsymbol{j}}+b\right)=\frac{1}{2}\\ \lim_{b\uparrow -r_{\boldsymbol{j}}}\lim_{\zeta\to\infty} C(r_{\boldsymbol{j}},\zeta,b) &= \lim_{b\uparrow -r_{\boldsymbol{j}}}\lim_{\zeta\to\infty}\Pr\left(\frac{\Delta^s y_{it}}{\Delta^p y_{it-s}-\zeta}\le r_{\boldsymbol{j}}+b\right)=0\\ \lim_{b\uparrow -r_{\boldsymbol{j}}}\lim_{\zeta\to\infty} D(r_{\boldsymbol{j}},\zeta,b) &= \lim_{b\uparrow -r_{\boldsymbol{j}}}\lim_{\zeta\to\infty}\Pr\left(\frac{\Delta^s y_{it}+\zeta}{\Delta^p y_{it-s}-\zeta}\le r_{\boldsymbol{j}}+b\right)=1\end{aligned} \tag{3.72}$$

and

$$\lim_{b\uparrow -r_{\boldsymbol{j}}} A(r_{\boldsymbol{j}},b) = \lim_{b\uparrow -r_{\boldsymbol{j}}}\frac{1}{2}+\frac{1}{\pi}\arctan\sigma^* b = \frac{1}{2}+\frac{1}{\pi}\arctan\sigma^*(-r_{\boldsymbol{j}}), \tag{3.73}$$

where the last limit follows from (3.63). We have that

$$\lim_{b\uparrow -r_{\boldsymbol{j}}} \lim_{\zeta\to\infty} V(r_{\boldsymbol{j}},\zeta,b,k,\epsilon_2) = \\ (1-\mathfrak{p}_B-\mathfrak{p}_C-\mathfrak{p}_D)\left(\frac{1}{2}+\frac{1}{\pi}\arctan\sigma^*(-r_{\boldsymbol{j}})\right)+\frac{1}{2}\mathfrak{p}_B+\mathfrak{p}_D \leq \frac{1}{2}, \tag{3.74}$$

which implies

$$2\frac{1-\mathfrak{p}_B-\mathfrak{p}_C-\mathfrak{p}_D}{\mathfrak{p}_C-\mathfrak{p}_D}\frac{1}{\pi}\arctan\sigma^*(-r_{\boldsymbol{j}}) \leq 1. \tag{3.75}$$

Replacing σ^* by (3.58), $\mathfrak{p}_B$, $\mathfrak{p}_C$, and $\mathfrak{p}_D$ by (3.67)–(3.69), and $r_{\boldsymbol{j}} = -(1-\alpha^s)/2$ completes the proof. □

Proof of Theorem 8. By the definition of influence function in (3.22)

$$\mathrm{IF}(\hat{r}_{\boldsymbol{j}}; Z^\alpha, Z^{2,\epsilon,\zeta,k}) = \left.\frac{\partial b(r_{\boldsymbol{j}},\zeta,\epsilon_1,k)}{\partial\epsilon_1}\right|_{\epsilon_1=0}, \tag{3.76}$$

where b denotes the bias of $\hat{r}_{\boldsymbol{j}}$. Given that $(1-\epsilon_2)^k = 1-\epsilon_1$, it holds

$$\frac{\partial b(r_{\boldsymbol{j}},\zeta,\epsilon_1,k)}{\partial\epsilon_1} = \frac{\partial b(r_{\boldsymbol{j}},\zeta,\epsilon_1,k)}{\partial\epsilon_2}\frac{\partial\epsilon_2}{\partial\epsilon_1} = \frac{\partial b(r_{\boldsymbol{j}},\zeta,\epsilon_1,k)}{\partial\epsilon_2}\frac{1}{k(1-\epsilon_2)^{k-1}}. \tag{3.77}$$

The derivative in (3.77) can obtained by applying the implicit function theorem to (3.66),

$$\left.\frac{\partial b(r_{\boldsymbol{j}},\zeta,\epsilon_1,k)}{\partial\epsilon_2}\right|_{\epsilon_1=0} = -\frac{\sum\limits_{j\in\{B,C,D\}}\mathfrak{p}'_j(0)j(r_{\boldsymbol{j}},\zeta,0)+\mathfrak{p}'_A(0)A(r_{\boldsymbol{j}},0)}{A'_b(r_{\boldsymbol{j}},0)}, \tag{3.78}$$

where $A'_b(r_{\boldsymbol{j}},0)$ is the same as in (3.62) and where $\mathfrak{p}'_j$, $j\in\{B,C,D\}$, denotes the derivative of $\mathfrak{p}_j$ in Equations (3.67)–(3.69) with respect to ϵ_2, that is

$$\mathfrak{p}'_B(0) := \left.\frac{\partial\mathfrak{p}_B(\epsilon_2;s,p,k)}{\partial\epsilon_2}\right|_{\epsilon_1=0} \tag{3.79}$$
$$= \min\{s,k\}+\max\Big\{0, s+k-\max\{s+p,k\}\Big\},$$
$$\mathfrak{p}'_C(0) = p+k-\max\{p,k\}+\max\Big\{0,\min\{s+p,k\}-s\Big\} \tag{3.80}$$
$$\mathfrak{p}'_D(0) = \max\Big\{0, s+\min\{p,k\}-\max\{s,k\}\Big\} \tag{3.81}$$

and

$$\mathfrak{p}'_A(0) = -[\mathfrak{p}'_B(0) + \mathfrak{p}'_C(0) + \mathfrak{p}'_D(0)]. \tag{3.82}$$

As in Section 3.A.2, $A(r_{\boldsymbol{j}}; 0) = 1/2$. Recall Equations (3.56)–(3.58). By Dhaene and Zhu (2009, Lemma 3),

$$B(r_{\boldsymbol{j}}; \zeta, 0) = \frac{1}{2} \tag{3.83}$$

$$C(r_{\boldsymbol{j}}; \zeta, 0) = \Phi\left(-\frac{r_{\boldsymbol{j}}\zeta}{\sigma_u}\right)\Phi\left(-\frac{\zeta}{\sigma_{\Delta^p}}\right) + \Phi\left(\frac{r_{\boldsymbol{j}}\zeta}{\sigma_u}\right)\Phi\left(\frac{\zeta}{\sigma_{\Delta^p}}\right) \tag{3.84}$$

$$D(r_{\boldsymbol{j}}; \zeta, 0) = \Phi\left(-\frac{(1+r_{\boldsymbol{j}})\zeta}{\sigma_u}\right)\Phi\left(-\frac{\zeta}{\sigma_{\Delta^p}}\right) + \Phi\left(\frac{(1+r_{\boldsymbol{j}})\zeta}{\sigma_u}\right)\Phi\left(\frac{\zeta}{\sigma_{\Delta^p}}\right), \tag{3.85}$$

where $\sigma_u := \sqrt{\mathrm{var}(u_{it\boldsymbol{j}})}$ and $\sigma_{\Delta^p} := \sqrt{\mathrm{var}(\Delta^p y_{it-s})}$ as in (3.48) and $r_{\boldsymbol{j}} = -(1-\alpha^s)/2$. Substituting (3.77)–(3.85) in (3.76) completes the proof. $\square$

Patch additive outlier contamination $Z^{3,\epsilon,\zeta,k}$

This case is a generalization of the $Z^{2,\epsilon,\zeta,k}$ contamination. The proof structure is very similar to the one in Sections 3.A.2 and 3.A.2, although the algebra is a bit more lengthy. As before, it is useful to derive first the bias of $\hat{r}_{\boldsymbol{j}}$ under the outlier contamination $Z^{3,\epsilon,\zeta,k}$ as defined in (3.14). Similarly to (3.45), this is given by $b := b(r_{\boldsymbol{j}}, \zeta, \epsilon_1, k)$ solving the equation

$$\begin{aligned}\Pr\left(\frac{\Delta^s y^{\epsilon,\zeta}_{it} - r_{\boldsymbol{j}}\Delta^p y^{\epsilon,\zeta}_{it-s}}{\Delta^p y^{\epsilon,\zeta}_{it-s}} \le b\right) &= \Pr\left(\frac{u_{it\boldsymbol{j}} + \Delta^s z^{\epsilon}_{it} - r_{\boldsymbol{j}}\Delta^p z^{\epsilon}_{it-s}}{\Delta^p y_{it-s} + \Delta^p z^{\epsilon}_{it-s}} \le b\right) \\ &= \mathfrak{p}_A A + \mathfrak{p}_B B + \mathfrak{p}_C C + \mathfrak{p}_D D + \mathfrak{p}_E E + \mathfrak{p}_F F + \mathfrak{p}_G G + \mathfrak{p}_H H + \mathfrak{p}_I I + \mathfrak{p}_J J = \frac{1}{2},\end{aligned} \tag{3.86}$$

where the notation is explained below. Note that the set Ω_{it} in (3.51) is different than it was for previous types of contaminations as now outliers can be either negative or positive.

Table 3.6: Configurations of patch outliers and their probabilities.

$(\lvert\zeta\rvert_1,\dots,0,\dots,0)'$	$(1-\epsilon_2)^{k+\min\{p,k\}}\cdot\epsilon_2\cdot\min\{s,k\}$
$(0,\dots,0,\dots\lvert\zeta\rvert_1)'$	$\epsilon_2\cdot\Big(p+k-\max\{p,k\}\Big)\cdot(1-\epsilon_2)^{k+\min\{s,k\}}$
$(0,\dots,\lvert\zeta\rvert_1,\dots,0)'$	$(1-\epsilon_2)^{2k}\cdot\epsilon_2\cdot\max\Big\{0,s+\min\{p,k\}-\max\{s,k\}\Big\}$
$(\lvert\zeta\rvert_1,\dots,\lvert\zeta\rvert_1,\dots,0)'$	$(1-\epsilon_2)^{k}\cdot\epsilon_2\cdot\max\Big\{0,\min\{s+p,k\}-s\Big\}$
$(\lvert\zeta\rvert_2,\dots,\lvert\zeta\rvert_1,\dots,0)'$	$(1-\epsilon_2)^{k}\cdot\epsilon_2^2\cdot\max\Big\{0,s+\min\{p,k\}-\max\{s,k\}\Big\}\cdot\min\{s,k\}$
$(0,\dots,\lvert\zeta\rvert_1,\dots,\lvert\zeta\rvert_1)'$	$\epsilon_2\cdot\max\Big\{0,s+k-\max\{s+p,k\}\Big\}\cdot(1-\epsilon_2)^{k}$
$(0,\dots,\lvert\zeta\rvert_2,\dots,\lvert\zeta\rvert_1)'$	$\epsilon_2^2\cdot\Big(p+k-\max\{p,k\}\Big)$ $\cdot\max\Big\{0,s+\min\{p,k\}-\max\{s,k\}\Big\}\cdot(1-\epsilon_2)^{k}$
$(\lvert\zeta\rvert_2,\dots,0,\dots,\lvert\zeta\rvert_1)'$	$\epsilon_2^2\cdot\Big(p+k-\max\{p,k\}\Big)\cdot(1-\epsilon_2)^{k}\cdot\min\{s,k\}$
$(\lvert\zeta\rvert_1,\dots,\lvert\zeta\rvert_1,\dots,\lvert\zeta\rvert_1)'$	$\epsilon_2\cdot\max\{0,k-s-p\}$
$(\lvert\zeta\rvert_2,\dots,\lvert\zeta\rvert_1,\dots,\lvert\zeta\rvert_1)'$	$\epsilon_2^2\cdot\max\{0,k-p\}\cdot\min\{s,k\}$
$(\lvert\zeta\rvert_2,\dots,\lvert\zeta\rvert_2,\dots,\lvert\zeta\rvert_1)'$	$\epsilon_2^2\cdot k\cdot\max\Big\{0,\min\{s+p,k\}-s\Big\}$
$(\lvert\zeta\rvert_3,\dots,\lvert\zeta\rvert_2,\dots,\lvert\zeta\rvert_1)'$	$\epsilon_2^3\cdot k\cdot\max\Big\{0,s+\min\{p,k\}-\max\{s,k\}\Big\}\cdot\min\{s,k\}$

By using the results in in Table 3.6, we have that

$$
\begin{aligned}
\mathfrak{p}_B &:= \Pr\left[\begin{pmatrix}\zeta\\0\\0\end{pmatrix}\cup\begin{pmatrix}-\zeta\\0\\0\end{pmatrix}\cup\begin{pmatrix}0\\\zeta\\\zeta\end{pmatrix}\cup\begin{pmatrix}0\\-\zeta\\-\zeta\end{pmatrix}\right]\\
&=\frac{1}{2}\Pr\begin{pmatrix}\zeta_1\\0\\0\end{pmatrix}+\frac{1}{2}\Pr\begin{pmatrix}-\zeta_1\\0\\0\end{pmatrix}+\frac{1}{4}\Pr\begin{pmatrix}0\\\zeta_2\\\zeta_1\end{pmatrix}+\frac{1}{4}\Pr\begin{pmatrix}0\\-\zeta_2\\-\zeta_1\end{pmatrix} \qquad (3.87)\\
&=\Pr\begin{pmatrix}\lvert\zeta\rvert_1\\0\\0\end{pmatrix}+\frac{1}{2}\Pr\begin{pmatrix}0\\\lvert\zeta\rvert_2\\\lvert\zeta\rvert_1\end{pmatrix},
\end{aligned}
$$

$$\begin{aligned}
\mathfrak{p}_C &:= \Pr\left[\begin{pmatrix}0\\0\\\zeta\end{pmatrix}\cup\begin{pmatrix}0\\0\\-\zeta\end{pmatrix}\cup\begin{pmatrix}\zeta\\\zeta\\0\end{pmatrix}\cup\begin{pmatrix}-\zeta\\-\zeta\\0\end{pmatrix}\right]\\
&= \Pr\begin{pmatrix}0\\0\\\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}0\\0\\-\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}\zeta_2\\\zeta_1\\0\end{pmatrix}+\Pr\begin{pmatrix}-\zeta_2\\-\zeta_1\\0\end{pmatrix} \qquad (3.88)\\
&= \Pr\begin{pmatrix}0\\0\\|\zeta|_1\end{pmatrix}+\frac{1}{2}\Pr\begin{pmatrix}|\zeta|_2\\|\zeta|_1\\0\end{pmatrix},
\end{aligned}$$

$$\begin{aligned}
\mathfrak{p}_D &:= \Pr\left[\begin{pmatrix}\zeta\\0\\\zeta\end{pmatrix}\cup\begin{pmatrix}-\zeta\\0\\-\zeta\end{pmatrix}\cup\begin{pmatrix}0\\\zeta\\0\end{pmatrix}\cup\begin{pmatrix}0\\-\zeta\\0\end{pmatrix}\right]\\
&= \Pr\begin{pmatrix}\zeta_2\\0\\\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}-\zeta_2\\0\\-\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}0\\\zeta_1\\0\end{pmatrix}+\Pr\begin{pmatrix}0\\-\zeta_1\\0\end{pmatrix} \qquad (3.89)\\
&= \frac{1}{2}\Pr\begin{pmatrix}|\zeta|_2\\0\\|\zeta|_1\end{pmatrix}+\Pr\begin{pmatrix}0\\|\zeta|_1\\0\end{pmatrix},
\end{aligned}$$

$$\begin{aligned}
\mathfrak{p}_E &:= \Pr\left[\begin{pmatrix}0\\-\zeta\\\zeta\end{pmatrix}\cup\begin{pmatrix}0\\\zeta\\-\zeta\end{pmatrix}\right]\\
&= \Pr\begin{pmatrix}0\\-\zeta_1\\\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}0\\-\zeta_2\\\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}0\\\zeta_1\\-\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}0\\\zeta_2\\-\zeta_1\end{pmatrix} \qquad (3.90)\\
&= \Pr\begin{pmatrix}0\\|\zeta|_1\\|\zeta|_1\end{pmatrix}+\frac{1}{2}\Pr\begin{pmatrix}0\\|\zeta|_2\\|\zeta|_1\end{pmatrix},
\end{aligned}$$

$$\begin{aligned}
\mathfrak{p}_F &:= \Pr\left[\begin{pmatrix}-\zeta\\0\\\zeta\end{pmatrix}\cup\begin{pmatrix}\zeta\\0\\-\zeta\end{pmatrix}\right] = \Pr\begin{pmatrix}-\zeta_2\\0\\\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}\zeta_2\\0\\-\zeta_1\end{pmatrix}\\
&= \frac{1}{2}\Pr\begin{pmatrix}|\zeta|_2\\0\\|\zeta|_1\end{pmatrix}, \qquad (3.91)
\end{aligned}$$

$$
\begin{aligned}
\mathfrak{p}_G &:= \Pr\left[\begin{pmatrix}\zeta\\-\zeta\\0\end{pmatrix}\cup\begin{pmatrix}-\zeta\\\zeta\\0\end{pmatrix}\right]\\
&= \Pr\begin{pmatrix}\zeta_1\\-\zeta_1\\0\end{pmatrix}+\Pr\begin{pmatrix}\zeta_2\\-\zeta_1\\0\end{pmatrix}+\Pr\begin{pmatrix}-\zeta_1\\\zeta_1\\0\end{pmatrix}+\Pr\begin{pmatrix}-\zeta_2\\\zeta_1\\0\end{pmatrix} \qquad (3.92)\\
&= \Pr\begin{pmatrix}|\zeta|_1\\|\zeta|_1\\0\end{pmatrix}+\frac{1}{2}\Pr\begin{pmatrix}|\zeta|_2\\|\zeta|_1\\0\end{pmatrix},
\end{aligned}
$$

$$
\begin{aligned}
\mathfrak{p}_H &:= \Pr\left[\begin{pmatrix}-\zeta\\-\zeta\\\zeta\end{pmatrix}\cup\begin{pmatrix}\zeta\\\zeta\\-\zeta\end{pmatrix}\right]\\
&= \Pr\begin{pmatrix}-\zeta_2\\-\zeta_1\\\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}-\zeta_3\\-\zeta_2\\\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}\zeta_2\\\zeta_1\\-\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}\zeta_3\\\zeta_2\\-\zeta_1\end{pmatrix} \qquad (3.93)\\
&= \frac{1}{2}\Pr\begin{pmatrix}|\zeta|_2\\|\zeta|_1\\|\zeta|_1\end{pmatrix}+\frac{1}{4}\Pr\begin{pmatrix}|\zeta|_3\\|\zeta|_2\\|\zeta|_1\end{pmatrix},
\end{aligned}
$$

$$
\begin{aligned}
\mathfrak{p}_I &:= \Pr\left[\begin{pmatrix}\zeta\\-\zeta\\\zeta\end{pmatrix}\cup\begin{pmatrix}-\zeta\\\zeta\\-\zeta\end{pmatrix}\right]\\
&= \Pr\begin{pmatrix}\zeta_1\\-\zeta_1\\\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}\zeta_2\\-\zeta_1\\\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}\zeta_2\\-\zeta_2\\\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}\zeta_3\\-\zeta_2\\\zeta_1\end{pmatrix}\\
&\quad+\Pr\begin{pmatrix}-\zeta_1\\\zeta_1\\-\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}-\zeta_2\\\zeta_1\\-\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}-\zeta_2\\\zeta_2\\-\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}-\zeta_3\\\zeta_2\\-\zeta_1\end{pmatrix} \qquad (3.94)\\
&= \Pr\begin{pmatrix}|\zeta|_1\\|\zeta|_1\\|\zeta|_1\end{pmatrix}+\frac{1}{2}\Pr\begin{pmatrix}|\zeta|_2\\|\zeta|_1\\|\zeta|_1\end{pmatrix}+\frac{1}{2}\Pr\begin{pmatrix}|\zeta|_2\\|\zeta|_2\\|\zeta|_1\end{pmatrix}+\frac{1}{4}\Pr\begin{pmatrix}|\zeta|_3\\|\zeta|_2\\|\zeta|_1\end{pmatrix},
\end{aligned}
$$

$$
\begin{aligned}
\mathfrak{p}_J &:= \Pr\left[\begin{pmatrix}\zeta\\-\zeta\\-\zeta\end{pmatrix}\cup\begin{pmatrix}-\zeta\\\zeta\\\zeta\end{pmatrix}\right]\\
&= \Pr\begin{pmatrix}\zeta_2\\-\zeta_2\\-\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}\zeta_3\\-\zeta_2\\-\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}-\zeta_2\\\zeta_2\\\zeta_1\end{pmatrix}+\Pr\begin{pmatrix}-\zeta_3\\\zeta_2\\\zeta_1\end{pmatrix} \qquad (3.95)\\
&= \frac{1}{2}\Pr\begin{pmatrix}|\zeta|_2\\|\zeta|_2\\|\zeta|_1\end{pmatrix}+\frac{1}{4}\Pr\begin{pmatrix}|\zeta|_3\\|\zeta|_2\\|\zeta|_1\end{pmatrix},
\end{aligned}
$$

and

$$
\mathfrak{p}_A = 1-\sum_{j\in\mathcal{I}\setminus\{A\}}\mathfrak{p}_j, \tag{3.96}
$$

where $\mathcal{I} := \{A,B,C,D,E,F,G,H,I,J\}$. Moreover,

$$
\begin{aligned}
A(r_{\boldsymbol{j}},b) &:= \Pr\left(\frac{u_{it\boldsymbol{j}}}{\Delta^p y_{it-s}}\le b\right)\\
B(r_{\boldsymbol{j}},\zeta,b) &:= \Pr\left(\frac{u_{it\boldsymbol{j}}+\zeta}{\Delta^p y_{it-s}}\le b\right)\\
C(r_{\boldsymbol{j}},\zeta,b) &:= \Pr\left(\frac{u_{it\boldsymbol{j}}+\zeta r_{\boldsymbol{j}}}{\Delta^p y_{it-s}-\zeta}\le b\right)\\
D(r_{\boldsymbol{j}},\zeta,b) &:= \Pr\left(\frac{u_{it\boldsymbol{j}}+\zeta(1+r_{\boldsymbol{j}})}{\Delta^p y_{it-s}-\zeta}\le b\right)\\
E(r_{\boldsymbol{j}},\zeta,b) &:= \Pr\left(\frac{u_{it\boldsymbol{j}}+\zeta(1+2r_{\boldsymbol{j}})}{\Delta^p y_{it-s}-2\zeta}\le b\right)\\
F(r_{\boldsymbol{j}},\zeta,b) &:= \Pr\left(\frac{u_{it\boldsymbol{j}}+\zeta(r_{\boldsymbol{j}}-1)}{\Delta^p y_{it-s}-\zeta}\le b\right)\\
G(r_{\boldsymbol{j}},\zeta,b) &:= \Pr\left(\frac{u_{it\boldsymbol{j}}+\zeta(2+r_{\boldsymbol{j}})}{\Delta^p y_{it-s}-\zeta}\le b\right)\\
H(r_{\boldsymbol{j}},\zeta,b) &:= \Pr\left(\frac{u_{it\boldsymbol{j}}+2\zeta r_{\boldsymbol{j}}}{\Delta^p y_{it-s}-2\zeta}\le b\right)\\
I(r_{\boldsymbol{j}},\zeta,b) &:= \Pr\left(\frac{u_{it\boldsymbol{j}}+2\zeta(1+r_{\boldsymbol{j}})}{\Delta^p y_{it-s}-2\zeta}\le b\right)\\
J(r_{\boldsymbol{j}},\zeta,b) &:= \Pr\left(\frac{u_{it\boldsymbol{j}}+2\zeta}{\Delta^p y_{it-s}}\le b\right),
\end{aligned} \tag{3.97}
$$

where the symmetry $L(k,l,b) = L(-k,-l,b)$ has been used, recall Equation (3.55).

Proof of Theorem 6. Denote the left hand side of (3.86) as $V(r_{\boldsymbol{j}},\zeta,b,k,\epsilon_2)$.

As $\zeta \to \infty$, $\hat{r}_{\boldsymbol{j}}$ breaks down to -1 if and only if

$$\lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} V(r_{\boldsymbol{j}}, \zeta, b, k, \epsilon_2) \leq \frac{1}{2}. \tag{3.98}$$

Since

$$\begin{aligned}
\lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} B(r_{\boldsymbol{j}}, \zeta, b) &= \lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} \Pr\left(\frac{\Delta^s y_{it} + \zeta}{\Delta^p y_{it-s}} \leq r_{\boldsymbol{j}} + b\right) = \frac{1}{2} \\
\lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} C(r_{\boldsymbol{j}}, \zeta, b) &= \lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} \Pr\left(\frac{\Delta^s y_{it}}{\Delta^p y_{it-s} - \zeta} \leq r_{\boldsymbol{j}} + b\right) = 0 \\
\lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} D(r_{\boldsymbol{j}}, \zeta, b) &= \lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} \Pr\left(\frac{\Delta^s y_{it} + \zeta}{\Delta^p y_{it-s} - \zeta} \leq r_{\boldsymbol{j}} + b\right) = 1 \\
\lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} E(r_{\boldsymbol{j}}, \zeta, b) &= \lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} \Pr\left(\frac{\Delta^s y_{it} + \zeta}{\Delta^p y_{it-s} - 2\zeta} \leq r_{\boldsymbol{j}} + b\right) = 0 \\
\lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} F(r_{\boldsymbol{j}}, \zeta, b) &= \lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} \Pr\left(\frac{\Delta^s y_{it} - \zeta}{\Delta^p y_{it-s} - \zeta} \leq r_{\boldsymbol{j}} + b\right) = 0 \\
\lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} G(r_{\boldsymbol{j}}, \zeta, b) &= \lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} \Pr\left(\frac{\Delta^s y_{it} + 2\zeta}{\Delta^p y_{it-s} - \zeta} \leq r_{\boldsymbol{j}} + b\right) = 1 \\
\lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} H(r_{\boldsymbol{j}}, \zeta, b) &= \lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} \Pr\left(\frac{\Delta^s y_{it}}{\Delta^p y_{it-s} - 2\zeta} \leq r_{\boldsymbol{j}} + b\right) = 0 \\
\lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} I(r_{\boldsymbol{j}}, \zeta, b) &= \lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} \Pr\left(\frac{\Delta^s y_{it} + 2\zeta}{\Delta^p y_{it-s} - 2\zeta} \leq r_{\boldsymbol{j}} + b\right) = 1 \\
\lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} J(r_{\boldsymbol{j}}, \zeta, b) &= \lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} \Pr\left(\frac{\Delta^s y_{it} + 2\zeta}{\Delta^p y_{it-s}} \leq r_{\boldsymbol{j}} + b\right) = \frac{1}{2}
\end{aligned} \tag{3.99}$$

and

$$\lim_{b \downarrow -1-r_{\boldsymbol{j}}} A(r_{\boldsymbol{j}}, b) = \lim_{b \downarrow -1-r_{\boldsymbol{j}}} \frac{1}{2} + \frac{1}{\pi} \arctan \sigma^* b = \frac{1}{2} + \frac{1}{\pi} \arctan \sigma^* (-1 - r_{\boldsymbol{j}}) \tag{3.100}$$

we have that

$$\lim_{b \downarrow -1-r_{\boldsymbol{j}}} \lim_{\zeta \to \infty} V(r_{\boldsymbol{j}}, \zeta, b, k, \epsilon_2) = \\
\left(1 - \sum_{j \in \mathcal{I} \setminus \{A\}} \mathfrak{p}_j\right) \left(\frac{1}{2} + \frac{1}{\pi} \arctan \sigma^* (-1 - r_{\boldsymbol{j}})\right) + \frac{1}{2}(\mathfrak{p}_B + \mathfrak{p}_J) + \mathfrak{p}_D + \mathfrak{p}_G + \mathfrak{p}_I \leq \frac{1}{2}, \tag{3.101}$$

which implies

$$2\frac{\mathfrak{p}_A}{\mathfrak{p}_C+\mathfrak{p}_E+\mathfrak{p}_F+\mathfrak{p}_H-(\mathfrak{p}_D+\mathfrak{p}_G+\mathfrak{p}_I)}\frac{1}{\pi}\arctan\sigma^*(-1-r_{\boldsymbol{j}})\leq 1. \quad (3.102)$$

□

Proof of Theorem 9. Denote

$$\mathfrak{p}'_j(0) := \left.\frac{\partial\mathfrak{p}_j(\epsilon_2;s,p,k)}{\partial\epsilon_2}\right|_{\epsilon_2=0} \qquad j\in\mathcal{I}:=\{A,B,C,D,E,F,G,H,I,J\},$$

where $\mathfrak{p}_j(\cdot)$, $j\in\mathcal{I}$, are defined in (3.87)–(3.96). Given that $(1-\epsilon_2)^k=1-\epsilon_1$, it holds

$$\begin{aligned}\mathrm{IF}(\hat{r}_{\boldsymbol{j}};Z^\alpha,Z^{3,\epsilon,\zeta,k}) &= \frac{\partial\,\mathrm{bias}(\hat{r}_{\boldsymbol{j}};Z^\alpha,Z^{3,\epsilon,\zeta,k})}{\partial\epsilon_1}=\frac{\partial b(r_{\boldsymbol{j}},\zeta,\epsilon_1,k)}{\partial\epsilon_1}=\frac{\partial b(r_{\boldsymbol{j}},\zeta,\epsilon_1,k)}{\partial\epsilon_2}\frac{\partial\epsilon_2}{\partial\epsilon_1}\\ &=\frac{\partial b(r_{\boldsymbol{j}},\zeta,\epsilon_1,k)}{\partial\epsilon_2}\frac{1}{k(1-\epsilon_2)^{k-1}}.\end{aligned} \quad (3.103)$$

Differentiating (3.86) and evaluating it at $\epsilon_1=0$ yields

$$\left.\frac{\partial b(r_{\boldsymbol{j}},\zeta,\epsilon_1,k)}{\partial\epsilon_1}\right|_{\epsilon_1=0}=-\frac{\sum_{j\in\mathcal{I}\setminus\{A\}}\mathfrak{p}'_j(0)j(r_{\boldsymbol{j}},\zeta,0)-A(r_{\boldsymbol{j}},0)\sum_{j\in\mathcal{I}\setminus\{A\}}\mathfrak{p}'_j(0)}{A'_b(r_{\boldsymbol{j}},0)}, \quad (3.104)$$

where $A'_b(r_{\boldsymbol{j}},0)$ is as in (3.62) and where (see results in Table 3.6)

$$\mathfrak{p}'_B(0)=\min\{s,k\} \quad (3.105)$$

$$\mathfrak{p}'_C(0)=p+k-\max\{p,k\} \quad (3.106)$$

$$\mathfrak{p}'_D(0)=\max\Big\{0,s+\min\{p,k\}-\max\{s,k\}\Big\} \quad (3.107)$$

$$\mathfrak{p}'_E(0)=\max\Big\{0,s+k-\max\{s+p,k\}\Big\} \quad (3.108)$$

$$\mathfrak{p}'_F(0)=0 \quad (3.109)$$

$$\mathfrak{p}'_G(0)=\max\Big\{0,\min\{s+p,k\}-s\Big\} \quad (3.110)$$

$$\mathfrak{p}'_H(0)=0 \quad (3.111)$$

$$\mathfrak{p}'_I(0)=\max\{0,k-s-p\} \quad (3.112)$$

$$\mathfrak{p}'_J(0)=0. \quad (3.113)$$

As in Section 3.A.2, $A(r_{\boldsymbol{j}};0) = 1/2$. Recall Equations (3.56)–(3.58). By Dhaene and Zhu (2009, Lemma 3),

$$B(r_{\boldsymbol{j}};\zeta,0) = \frac{1}{2} \tag{3.114}$$

$$C(r_{\boldsymbol{j}};\zeta,0) = \Phi\left(-\frac{r_{\boldsymbol{j}}\zeta}{\sigma_u}\right)\Phi\left(-\frac{\zeta}{\sigma_{\Delta^p}}\right) + \Phi\left(\frac{r_{\boldsymbol{j}}\zeta}{\sigma_u}\right)\Phi\left(\frac{\zeta}{\sigma_{\Delta^p}}\right) \tag{3.115}$$

$$D(r_{\boldsymbol{j}};\zeta,0) = \Phi\left(-\frac{(1+r_{\boldsymbol{j}})\zeta}{\sigma_u}\right)\Phi\left(-\frac{\zeta}{\sigma_{\Delta^p}}\right) + \Phi\left(\frac{(1+r_{\boldsymbol{j}})\zeta}{\sigma_u}\right)\Phi\left(\frac{\zeta}{\sigma_{\Delta^p}}\right) \tag{3.116}$$

$$E(r_{\boldsymbol{j}};\zeta,0) = \Phi\left(-\frac{(1+2r_{\boldsymbol{j}})\zeta}{\sigma_u}\right)\Phi\left(-\frac{2\zeta}{\sigma_{\Delta^p}}\right) + \Phi\left(\frac{(1+2r_{\boldsymbol{j}})\zeta}{\sigma_u}\right)\Phi\left(\frac{2\zeta}{\sigma_{\Delta^p}}\right) \tag{3.117}$$

$$G(r_{\boldsymbol{j}};\zeta,0) = \Phi\left(-\frac{(2+r_{\boldsymbol{j}})\zeta}{\sigma_u}\right)\Phi\left(-\frac{\zeta}{\sigma_{\Delta^p}}\right) + \Phi\left(\frac{(2+r_{\boldsymbol{j}})\zeta}{\sigma_u}\right)\Phi\left(\frac{\zeta}{\sigma_{\Delta^p}}\right) \tag{3.118}$$

$$I(r_{\boldsymbol{j}};\zeta,0) = \Phi\left(-\frac{2(1+r_{\boldsymbol{j}})\zeta}{\sigma_u}\right)\Phi\left(-\frac{2\zeta}{\sigma_{\Delta^p}}\right) + \Phi\left(\frac{2(1+r_{\boldsymbol{j}})\zeta}{\sigma_u}\right)\Phi\left(\frac{2\zeta}{\sigma_{\Delta^p}}\right), \tag{3.119}$$

where $\sigma_u := \sqrt{\mathrm{var}(u_{it\boldsymbol{j}})}$ and $\sigma_{\Delta^p} := \sqrt{\mathrm{var}(\Delta^p y_{it-s})}$. Replacing (3.104)–(3.119) in (3.103) completes the proof. □

General results

Theorem 10. The result follows directly from the definition of the breakdown point. The GMM estimator $\hat{\alpha}_n$ minimizes $\sum_{\boldsymbol{j}\in\mathcal{J}} A_{\boldsymbol{jj}} g^2_{n\boldsymbol{j}}(\alpha)$. As this objective function depends on the data only by means of the estimates $\hat{r}_{n\boldsymbol{j}}$, the estimator $\hat{\alpha}_n$ breaks down, that is, becomes fully independent of data, if and only if all estimates $\hat{r}_{n\boldsymbol{j}}$ break down. In other words, as long as $\hat{r}_{n\boldsymbol{j}}$ has a non-degenerate distribution for some $\boldsymbol{j}'$, $\hat{\alpha}_n$ will have a non-degenerate distribution too because all other values $\hat{r}_{n\boldsymbol{j}}, \boldsymbol{j} \neq \boldsymbol{j}'$, break down towards a finite c. □

Proof of Theorem 11. The proof follows directly from Equations (3.37) and (3.43). The estimator $\hat{\alpha}_n$ is defined by the solution of the sample analogs of equations (3.9), which are deterministic functions of $\hat{r}_{n\boldsymbol{j}}$. Thus the influence function of $\hat{\alpha}_n$ is fully determined by the influence functions of each $\hat{r}_{n\boldsymbol{j}}$, that is

$$\mathrm{IF}(\hat{\alpha}; Z^{\alpha}, Z^{\mathfrak{J},\epsilon,\zeta}) = -(\boldsymbol{d}'\boldsymbol{A}\boldsymbol{d})^{-1}\boldsymbol{d}'\boldsymbol{A}\boldsymbol{\psi}, \tag{3.120}$$

where $\boldsymbol{\psi} := \big(\mathrm{IF}(\hat{r}_{\boldsymbol{j}}; Z^{\alpha}, Z^{\mathfrak{I},\epsilon,\zeta})\big)_{\boldsymbol{j}\in\mathcal{J}_o}$ is a $\#\mathcal{J}_o \times 1$ vector whose elements $\mathrm{IF}(\hat{r}_{\boldsymbol{j}}; Z^{\alpha}, Z^{\mathfrak{I},\epsilon,\zeta})$, $\boldsymbol{j} \in \mathcal{J}_o$, are derived for each considered data contamination $Z^{1,\epsilon,\zeta}$, $Z^{2,\epsilon,\zeta,k}$, and $Z^{3,\epsilon,\zeta,k}$ in Theorem 7, 8, and 9 respectively. $\square$

Chapter 4

Pairwise difference estimation of dynamic panel data models[1]

4.1 Introduction

The estimation of the dynamic linear panel data model with fixed effects has been extensively studied in last decades. It is well known that the least square dummy variable (LSDV) estimator is inconsistent when applied to dynamic panels with a small fixed number of time periods (Nickell, 1981). As a consequence, the majority of research has focused on the generalized method of moments (GMM) procedures and estimation methods based on instrumental variable (IV) methods (e.g., Anderson and Hsiao, 1982; Holtz-Eakin et al., 1988; Arellano and Bond, 1991; Arellano and Bover, 1995; Blundell and Bond, 1998). However, some of these estimators have been found to suffer heavily from various sources of finite sample bias. Models in first differences with instruments in levels (Arellano and Bond, 1991) can be substantially biased, in particular when the autoregressive parameter is close to unity. Models in levels with instruments in first differences as the one proposed by Arellano and Bover (1995) and implemented by Blundell and Bond (1998) are specifically designed for persistent series and rely crucially on the stationarity assumption (see Hahn, 1999). As an alternative approach, bias-reduction methods for the LSDV and maximum likelihood estimators have been proposed, see Kiviet (1995), Hahn and Kuersteiner (2002), Bun and Carree (2005), and Gouriéroux et al. (2010).

[1]This chapter is based on Aquaro M., and P. Čížek (2012), Pairwise difference estimation of dynamic panel data models, *Working Paper*.

To improve the estimation when the autoregressive parameter is close to 1 or -1, Hahn et al. (2007) suggest to employ the longest difference (LD) of the model, that is, the differences between the last and the first observation for each individual. Contrary to Blundell and Bond (1998), Hahn et al. (2007) derive valid moment conditions without imposing the stationarity assumption. Using the local-to-unity asymptotics for the autoregressive parameter, the moment conditions defining the LD estimator of Hahn et al. (2007) are chosen from the asymptotically relevant moment conditions in order to minimize the estimator's bias. Additionally, to circumvent the nonlinearity of the proposed moment conditions, the instruments – being regression residuals – are estimated using an initial consistent estimate.

Although the LD estimator provides a method with a small finite-sample bias without assuming stationarity, there are two important deficiencies of the method from the practical point of view. First, by using the longest difference of the panel data, the differenced data always contain only one observation for each individual unit irrespective of the number of time periods. Next, a practically applicable asymptotic distribution and variance of the LD estimator is not provided. This is especially important due to the reliance of the LD moment conditions on initially estimated instruments and thus on the properties of the initial estimator.

To rectify these problems and make LD a practically relevant alternative to the GMM estimators such as the one by Blundell and Bond (1998), two steps are necessary. First, we propose to extend the LD estimator by taking more longer differences than just the longest one and show that new estimators have smaller variances while keeping the bias properties almost unchanged. The proposed estimators improve upon the original LD especially for small values autoregressive parameter or larger numbers of time periods. Second, we derive the practically applicable asymptotic-distribution expression for a general class of long-difference estimators — including the original LD estimator — under strong instrument asymptotics. Practical choices and recommendations for the GMM weight matrix are extensively discussed as well. Finally, the theoretical findings are confirmed in finite samples by means of simulation studies.

The rest of the paper is organized as follows. In Section 4.2, we introduce the dynamic panel data model and the LD estimator. The new estimators are presented in Section 4.3, where we also study their bias properties. The asymptotic distribution for a finite number of time periods is derived in Section 4.4. Further, Section 4.5 contains the results of the Monte Carlo experiments. Finally, Section 4.6 concludes. The proofs are provided in the Appendix.

4.2 Long difference estimation of dynamic panels

For a fixed $T \geq 3$ and $n \in \mathbb{N}$, consider the simple dynamic panel data model

$$y_{it} = \alpha y_{i(t-1)} + \eta_i + \varepsilon_{it} \qquad (t = 1, \dots, T; \quad i = 1, \dots, n), \tag{4.1}$$

together with the assumption

$$\mathrm{E}(\varepsilon_{it} | y_{i(t-1)}, \dots, y_{i0}, \eta_i) = 0 \qquad (t = 1, \dots, T), \tag{4.2}$$

where the response variable y_{it} depends on its lagged value $y_{i(t-1)}$ through the unknown autoregressive parameter α, $|\alpha| < 1$, on the unobserved individual fixed effect η_i, and on an idiosyncratic error ε_{it}. Model (4.1) will be used to describe the estimation concepts, but it can and will be further generalized by including additional explanatory variables; see Section 4.4.

As the individual effects η_i are not observed, several filtering data transformations have been used in the literature. Many of those rely on the sth difference transformation generically defined as $\Delta^s v_t = v_t - v_{t-s}$ (see Aquaro and Čížek, 2010). More specifically, subtracting (4.1) at time $t - s$ from its level at time t yields

$$\Delta^s y_{it} = \alpha \Delta^s y_{i(t-1)} + \Delta^s \varepsilon_{it} \qquad (t = s+1, \dots, T; \quad i = 1, \dots, n), \tag{4.3}$$

where the order of the difference s generally ranges from 1 to $T - 1$: the Arellano and Bond (1991) use $s = 1$, whereas Hahn et al. (2007) employ $s = T - 1$. Aggregating across all time periods and using a vector notation, a more compact notation is $\boldsymbol{D}_s \boldsymbol{y}_i = \alpha \boldsymbol{D}_s \boldsymbol{y}_{i(-1)} + \boldsymbol{D}_s \boldsymbol{\varepsilon}_i$, where $\boldsymbol{D}_s$ is the $(T - s) \times T$ sth difference-operator matrix ($\boldsymbol{D}_s = (\boldsymbol{I}_{T-s}, \boldsymbol{0}) - (\boldsymbol{0}, \boldsymbol{I}_{T-s})$), $\boldsymbol{y}_i = (y_{i1}, \dots, y_{iT})'$, $\boldsymbol{y}_{i(-1)} = (y_{i0}, \dots, y_{i(T-1)})'$, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{iT})'$.

Hahn et al. (2007) propose to estimate α in (4.1) by using the long $(T - 1)$th difference technique of Griliches and Hausman (1986). The model (4.3) then becomes

$$\Delta^{T-1} y_{iT} = y_{iT} - y_{i1} = \alpha(y_{i(T-1)} - y_{i0}) + \varepsilon_{iT} - \varepsilon_{i1} = \alpha \Delta^{T-1} y_{i(T-1)} + \Delta^{T-1} \varepsilon_{iT}$$

(4.4)

for $i = 1, \ldots, n$. Under the assumption in (4.2), the long difference (LD) estimator itself is based on the following $T-1$ moment conditions, $T \geq 3$:

$$\mathrm{E}[y_{i0}\Delta^{T-1}\varepsilon_{iT}] = \mathrm{E}[y_{i0}(\varepsilon_{iT} - \varepsilon_{i1})] = 0, \tag{4.5a}$$

$$\mathrm{E}[u_{ir}\Delta^{T-1}\varepsilon_{iT}] = \mathrm{E}\left[u_{ir}(\varepsilon_{iT} - \varepsilon_{i1})\right] = 0 \qquad (r = 2, \ldots, T-1), \tag{4.5b}$$

where $u_{ir} = y_{ir} - \alpha y_{i(r-1)} = \eta_i + \varepsilon_{ir}$ (if $T = 2$, only moment condition (4.5a) makes sense and LD corresponds to the Arellano and Bond (1991) estimator). The operational moment conditions are then obtained by substituting for $\Delta^{T-1}\varepsilon_{iT}$ from (4.4). The moment conditions however contain also unobservable residuals u_{ir}. To produce moment conditions linear in α, a preliminary consistent estimator $\hat{\alpha}_n^0$ of α has to be used to compute and substitute estimates $\hat{u}_{ir} = y_{ir} - \hat{\alpha}_n^0 y_{i(r-1)}$ into (4.5b). Hahn et al. (2007) studied the GMM estimator based on the moment conditions (4.5) under the local-to-unity asymptotics, that is, assuming $\alpha_n \to 1$ as $n \to +\infty$, to confirm that these moment conditions do not rely on weak instruments in this limit case.

4.3 A class of long difference estimators

Apart for a more complicated asymptotic distribution of the LD estimator caused by estimating some instruments (see Section 4.4), an important disadvantage of the LD estimator is that, independently of the number T of time periods available, only a single observation per individual can be used after that data have been transformed by the long difference (4.4). This drawback is particularly problematic for data with a larger number T of time periods and a small or moderately large number n of individuals. As a remedy, we propose to extend the LD estimator by using multiple pairwise differences.

4.3.1 Pairwise-difference long-difference estimator

Let S denote the shortest difference considered in estimation, $2 \leq S \leq T-1$. To estimate α in (4.1), we propose to use the moment conditions of the LD estimator obtained by taking the longest difference Δ^{T-1}, the second longest differences Δ^{T-2}, and so on down to the differences Δ^{S}. This leads to $T^* = (T-S)(T-S+1)/2$ differenced equations consisting of all pairwise differences of lengths greater or equal to S:

$$
\begin{aligned}
s = T-1: \quad & \Delta^{T-1} y_{iT} = \alpha \Delta^{T-1} y_{i(T-1)} + \Delta^{T-1} \varepsilon_{iT}, && (4.6)\\
s = T-2: \quad & \Delta^{T-2} y_{iT} = \alpha \Delta^{T-2} y_{i(T-1)} + \Delta^{T-2} \varepsilon_{iT}, && (4.7)\\
& \Delta^{T-2} y_{i(T-1)} = \alpha \Delta^{T-2} y_{i(T-2)} + \Delta^{T-2} \varepsilon_{i(T-1)}, && (4.8)\\
\vdots \qquad & \qquad \vdots \\
s = S \quad : \quad & \Delta^{S} y_{iT} = \alpha \Delta^{S} y_{i(T-1)} + \Delta^{S} \varepsilon_{iT}, && (4.9)\\
& \qquad \vdots \\
& \Delta^{S} y_{i(S+1)} = \alpha \Delta^{S} y_{iS} + \Delta^{S} \varepsilon_{i(S+1)}. && (4.10)
\end{aligned}
$$

Obviously, if $S = T-1$, this system of differenced equations reduces to the original LD equation (4.4). In general, the shortest difference $S < T-1$ should be chosen so that the number of equations $T^* \leq (T-1)$, which implies that $S > T - \sqrt{2T}$. If $T^* > (T-1)$, some of the moment equations implied by the model could be written as a linear combination of the other ones and would not contribute new information to the system (e.g., in the extreme case of $S = 1$, any sth difference equation could be written as a sum of the consecutive first-differenced equations). This observations is a special case of the equivalence statement in Arellano and Bover (1995).

Under the assumption in (4.2), using the instruments (4.5) for each of the above stated equations, $s = S, \ldots, T-1$, leads to the set of the following moment conditions defining the infeasible pairwise-difference long-difference (PD-LD) estimator:

$$
\begin{aligned}
\mathrm{E}[y_{i(t-s-1)} \Delta^s \varepsilon_{it}] &= \mathrm{E}[y_{i(t-s-1)}(\varepsilon_{it} - \varepsilon_{i(t-s)})] = 0, && (4.11)\\
\mathrm{E}[u_{i(t-1)} \Delta^s \varepsilon_{it}] &= \mathrm{E}[u_{i(t-1)}(\varepsilon_{it} - \varepsilon_{i(t-s)})] = 0, && (4.12)\\
&\vdots \\
\mathrm{E}[u_{i(t-s+1)} \Delta^s \varepsilon_{it}] &= \mathrm{E}[u_{i(t-s+1)}(\varepsilon_{it} - \varepsilon_{i(t-s)})] = 0, && (4.13)
\end{aligned}
$$

where $t = s+1, \ldots, T$ and $s = S, \ldots, T-1$.

To express the PD-LD estimator explicitly as a GMM estimator, let us first write the moment conditions (4.11)–(4.13) for a single equation in a more compact form as

$$
\mathrm{E}(\boldsymbol{z}_{its} \Delta^s \varepsilon_{it}) = \mathbf{0} \qquad (t = s+1, \ldots, T; \quad s = S, \ldots, T-1), \tag{4.14}
$$

where $\boldsymbol{z}_{its}$ is a $s \times 1$ vector $\boldsymbol{z}_{its} = (y_{i(t-s-1)}, u_{i(t-1)}, \ldots, u_{i(t-s+1)})'$. Further-more, writing the equations (4.6)–(4.10) in the matrix form, the PD-LD estimator is based on the following differenced equations

$$\boldsymbol{D}\boldsymbol{y}_i = \alpha \boldsymbol{D}\boldsymbol{y}_{i(-1)} + \boldsymbol{D}\boldsymbol{\varepsilon}_i \qquad (i = 1, \ldots, n), \tag{4.15}$$

where $\boldsymbol{D}$ is a $T^* \times T$ partitioned matrix, $\boldsymbol{D} = (\boldsymbol{D}'_S, \ldots, \boldsymbol{D}'_{T-1})'$. Hence, the complete set of the PD-LD moment conditions can be expressed in the matrix form as $\mathrm{E}[\boldsymbol{Z}'_i \boldsymbol{D}\boldsymbol{\varepsilon}_i] = 0$, where $\boldsymbol{Z}_i = \mathrm{diag}(\{\boldsymbol{z}'_{its}\}_{(t,s)\in\mathcal{T}})$, $\mathcal{T} = \{(t,s) : t = s+1, \ldots, T; s = S, \ldots, T-1\}$, denotes a block-diagonal matrix with T^* blocks $\boldsymbol{z}'_{its}$ indexed by $t = s+1, \ldots, T$ and $s = S, \ldots, T-1$.

As $\boldsymbol{z}_{its}$ in (4.14) is only partially observable, this PD-LD estimator is infeasible and a preliminary consistent estimator is needed to construct instruments. Let $\hat{\alpha}_n^0$ denote a preliminary consistent estimator of α (e.g., the Arellano-Bond estimator). The feasible instruments to be used in PD-LD are then $\hat{\boldsymbol{z}}_{its} = (y_{i(t-1-s)}, \hat{u}_{i(t-1)}, \ldots, \hat{u}_{i(t+1-s)})'$, where $\hat{u}_{ir} = y_{ir} - \hat{\alpha}_n^0 y_{i(r-1)}$; the corresponding feasible matrix representation will be denoted $\hat{\boldsymbol{Z}}_i$. Denoting the inverse weight matrix $\hat{\boldsymbol{V}}_n$, the feasible PD-LD estimator – being a standard GMM estimator with linear moment conditions – can be then defined as follows:

$$\hat{\alpha}_n^{\text{PD-LD}} = \left(\boldsymbol{y}^{*\prime}_{-1} \hat{\boldsymbol{Z}} \hat{\boldsymbol{V}}_n^{-1} \hat{\boldsymbol{Z}}' \boldsymbol{y}^*_{-1}\right)^{-1} \boldsymbol{y}^{*\prime}_{-1} \hat{\boldsymbol{Z}} \hat{\boldsymbol{V}}_n^{-1} \hat{\boldsymbol{Z}}' \boldsymbol{y}^*, \tag{4.16}$$

where $\boldsymbol{y}^* = ([\boldsymbol{D}\boldsymbol{y}_1]', \ldots, [\boldsymbol{D}\boldsymbol{y}_n]')'$ and $\boldsymbol{y}^*_{-1} = ([\boldsymbol{D}\boldsymbol{y}_{1(-1)}]', \ldots, [\boldsymbol{D}\boldsymbol{y}_{n(-1)}]')'$ are the differenced variables and $\hat{\boldsymbol{Z}} = (\hat{\boldsymbol{Z}}'_1, \ldots, \hat{\boldsymbol{Z}}'_n)'$ is an estimate of $\boldsymbol{Z} = (\boldsymbol{Z}'_1, \ldots, \boldsymbol{Z}'_n)'$. In other words, $\hat{\boldsymbol{Z}}\boldsymbol{y}^* = \sum_{i=1}^n \hat{\boldsymbol{Z}}_i \boldsymbol{y}^*_i$ and $\hat{\boldsymbol{Z}}\boldsymbol{y}^*_{-1} = \sum_{i=1}^n \hat{\boldsymbol{Z}}_i \boldsymbol{y}^*_{i(-1)}$, where $\boldsymbol{y}^*_i = \boldsymbol{D}\boldsymbol{y}_i$ and $\boldsymbol{y}^*_{i(-1)} = \boldsymbol{D}\boldsymbol{y}_{i(-1)}$.

By using $\hat{\alpha}_n^{\text{PD-LD}}$ to re-estimate $\hat{\boldsymbol{z}}_{its}$, one can iterate to another LD estimator, which will be referred to as PD-LD1. Eventually, the procedure can be further iterated, yielding PD-LD2, PD-LD3, and so on.

4.3.2 Mixed-distance long-difference estimator

Loosely speaking, the idea of taking the longest differences is based on the fact that moment conditions based on such a data transformation do not become weak when α approaches one (Hahn et al., 2007). Considering its pairwise-difference extensions, there are other alternative choices of $T-1$ differenced equations than just (4.6)–(4.10). For instance, one could make use of all possible pairwise differences from the shortest one $S = 2$ to the longest one $T-1$ and take only one equation for each $s, S \le s \le T-1$,

Table 4.1: Asymptotic bias and variance

Estimator	Limited # of instruments	Unlimited # of instruments
AB	$O(n^{-1}T^{-1})$	$O(n^{-1})$
LD	$O(n^{-1}\|\alpha\|^{T})$	$O(n^{-1}T\|\alpha\|^{T})$
MD-LD	$O(n^{-1}T^{-1})$	$O(n^{-1}T^{-1})$
PD-LD	$O(n^{-1}\|\alpha\|^{T-\sqrt{2T}})$	$O(n^{-1}T\|\alpha\|^{T-\sqrt{2T}})$

Note: AB refers to the Arellano-Bond estimator with the model in forward orthogonal deviations derived by Bun and Kiviet (2006).

in order to fulfill the condition that the number of employed equations $T^* \leq T-1$. As a reference example, let $\hat{\alpha}_n^{\text{MD-LD}}$ be the GMM estimator based on the following $(T-2)(T-1)/2$ moment conditions:

$$\mathrm{E}(\boldsymbol{z}_{its}\Delta^s\varepsilon_{it}) = \mathbf{0} \qquad (t = s+1; \quad s = 2, \ldots, T-1), \tag{4.17}$$

where the moment conditions are derived under the assumption in (4.2). This estimator will be referred to as the mixed-difference long-difference estimator (MD-LD) as it relies both on short and long differences. It will be shown that including shorter differences affects unfavourably the bias of the estimator, at least for a large T.

4.3.3 Finite sample bias

To compare the LD, PD-LD, and MD-LD estimators, we first derive the leading terms of their finite sample biases. For the sake of simplicity, we compare the methods in the infeasible setting. In this section, we also use the two-stage least squares weight matrix for all estimators (e.g., the inverse weight matrix for the PD-LD in (4.16) will be $\hat{\boldsymbol{V}}_n = \boldsymbol{V}_n = \sum_i \boldsymbol{Z}_i'\boldsymbol{Z}_i$ instead of a general one), which happens to be the optimal weight matrix for the infeasible LD estimator. Other (asymptotic) properties are studied under more general assumptions in Section 4.4.

To derive the biases of the long-difference estimators, we need to impose the following conditions (using one high-level assumption for simplicity):

B.1 For all $i \in \mathbb{N}$ and $t \in \mathbb{N}$, idiosyncratic shocks ε_{it} are mutually independent, have finite second moments, and $\mathrm{E}(\varepsilon_{it}|y_{i(t-1)}, \ldots, y_{i0}, \eta_i) = 0$ and $\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon_{it})$.

B.2 Let individual effects η_i be independently distributed across individuals with finite second moments.

B.3 Denoting $q_{nT^*} = \boldsymbol{y}_{-1}^{*\prime} \boldsymbol{Z} \boldsymbol{V}_n^{-1} \boldsymbol{Z}' \boldsymbol{y}_{-1}^*$, let p-lim $q_{nT^*}/(nT^*) = \bar{q} > 0$ for $nT^* \to \infty$, where $T^* \leq T-1$; in particular, $T^* = 1$ in the case of the LD estimator.

Theorem 12. *Let y_{it} be generated by* (4.1) *with $0 < |\alpha| < 1$ and let $\hat{\alpha}_{nT}^{LD}$, $\hat{\alpha}_{nT}^{PD\text{-}LD}$, and $\hat{\alpha}_{nT}^{MD\text{-}LD}$ be the infeasible two-stage least squares estimators based on moment conditions* (4.5), (4.14), *and* (4.17), *respectively. Additionally, suppose that Assumptions B.1–B.3 hold. When all possible instruments are included, the finite-sample biases of each estimator in the LD class are given by*

$$B_{LD} = O((nT_{LD}^*)^{-1}) \cdot \left(-\frac{\sigma_\varepsilon^2}{\alpha^2}(T-1)\alpha^T \right) = O(n^{-1}T|\alpha|^T), \quad (4.18)$$

$$B_{MD\text{-}LD} = O((nT_{MD}^*)^{-1}) \cdot \left[-\frac{\sigma_\varepsilon^2}{\alpha} \left(\frac{\alpha^2 - \alpha^T}{(1-\alpha)^2} - \frac{(T-1)\alpha^T - \alpha^2}{1-\alpha} \right) \right] = O(n^{-1}T^{-1}), \quad (4.19)$$

$$\begin{aligned} B_{PD\text{-}LD} = {} & O((nT_{PD}^*)^{-1}) \cdot \left\{ -\frac{\sigma_\varepsilon^2}{\alpha} \left[T \left(\frac{\alpha^S - \alpha^T}{(1-\alpha)^2} + \frac{(S-1)\alpha^S - (T-1)\alpha^T}{1-\alpha} \right) \right. \right. \\ & \left. \left. - 2\frac{\alpha^S - \alpha^T}{(1-\alpha)^3} + \frac{[2T-3]\alpha^T - [2S-3]\alpha^S}{(1-\alpha)^2} + \frac{(T-1)^2\alpha^T - (S-1)^2\alpha^S}{1-\alpha} \right] \right\} \\ = {} & O(n^{-1}T|\alpha|^{T-\sqrt{2T}}), \end{aligned} \quad (4.20)$$

where the leading terms in bounds $O(\cdot)$ are determined for $n \to \infty$ or $T \to \infty$ and $T_{LD}^* = 1$ in the case of LD, $T_{MD}^* = T-2$ in the case of ML-LD, and $S = \lceil T - \sqrt{2T} \rceil$ and $T_{PD}^* = (T-S)(T-S+1)/2$ in the case of PD-LD.

Similarly, when the number of instruments used for each moment equation is limited to be at most $\bar{m} \in \mathbb{N}$, the finite-sample biases are bounded by

$$|B_{LD}| \leq O((nT_{LD}^*)^{-1}) \cdot \sigma_\varepsilon^2 \bar{m} \alpha^{T-2} = O(n^{-1}|\alpha|^T), \quad (4.21)$$

$$|B_{MD-LD}| \leq O((nT_{MD}^*)^{-1}) \cdot \frac{\sigma_\varepsilon^2 \bar{m}}{|\alpha|} \cdot \frac{|\alpha|^2 - |\alpha|^T}{1-|\alpha|} = O(n^{-1}T^{-1}), \quad (4.22)$$

$$\begin{aligned} |B_{PD-LD}| \leq {} & O((nT_{PD}^*)^{-1}) \times \quad (4.23) \\ & \times \frac{\sigma_\varepsilon^2 \bar{m}}{|\alpha|} \left(T\frac{|\alpha|^S - |\alpha|^T}{1-|\alpha|} + \frac{|\alpha|^S - |\alpha|^T}{(1-|\alpha|)^2} + \frac{|(S-1)|\alpha|^S - (T-1)|\alpha|^T|}{1-|\alpha|} \right) \\ = {} & O(n^{-1}|\alpha|^{T-\sqrt{2T}}). \end{aligned} \quad (4.24)$$

Results concerning the leading terms are summarized in Table 4.1 (the order of bias of the Arellano-Bond estimator as derived in Bun and Kiviet (2006) is also reported). In general, the orders of biases are smaller when the number of employed instruments is limited, but the ranking of methods is not affected by the number of instruments. Taking into account (4.19), the infeasible LD and PD-LD methods exhibit the smallest biases, especially if T is large or α is small. The relatively small increase in bias of PD-LD relative to LD is substantially compensated by the fact that PD-LD uses $nT^* \approx n(T-1)$ observations compared to n observations used by LD (see also Section 4.4), which will complement the bias properties of PD-LD by a smaller variance of estimates compared to LD.

4.4 Asymptotic distribution

4.4.1 Asymptotic normality

In this section, the asymptotic distribution for the class of the long-difference estimators is derived. Although the asymptotic distribution of the LD estimator is derived in Hahn et al. (2007), it is given there only for the limit case of $\alpha \to 1$, without any exogenous variables, and in a form difficult for practical use.

For deriving the asymptotic distribution of different LD estimators, it is useful to generalize and derive this result for a model with exogenous variables. Let $\boldsymbol{x}_{it} = (x_{it1}, \ldots, x_{itK})'$ be a set of K exogenous or predetermined variables. Assume $T \geq 3$ is fixed and y_{it} follows

$$\begin{aligned} y_{it} &= \alpha y_{it-1} + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \eta_i + \varepsilon_{it}, \\ &= \boldsymbol{w}_{it}'\boldsymbol{\theta} + \eta_i + \varepsilon_{it}, \end{aligned} \qquad (t = 1, \ldots, T; \quad i = 1, \ldots, n), \quad (4.25)$$

where $\boldsymbol{w}_{it} = (w_{itk})_{k=1}^{K+1} = (y_{i(t-1)}, \boldsymbol{x}_{it}')'$ and the parameter of interest is $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}')'$ with the true value $\boldsymbol{\theta}^0$. Let $\boldsymbol{W}_i$ denote the $T \times (K+1)$ matrix $\boldsymbol{W}_i = (\boldsymbol{w}_{i1}, \ldots, \boldsymbol{w}_{iT})'$. The assumptions concerning the data-generating process (4.25), which is allowed to be heterogeneous across individuals i, follow.

A.1 Let $\{\boldsymbol{y}_i, \boldsymbol{W}_i, \eta_i\}_{i=1}^n$ be a sequence of independently distributed random vectors with uniformly bounded finite $(2+\delta)$th moments for some $\delta > 0$.

A.2 For all i and t, $\mathrm{E}(\varepsilon_{it} | \boldsymbol{w}_{it}, \ldots, \boldsymbol{w}_{i1}, \eta_i) = 0$.

Next, the initial estimator $\hat{\boldsymbol{\theta}}_n^0$ will be assumed to be a consistent GMM estimator of $\boldsymbol{\theta}^0$ based on moment conditions $\mathrm{E}[\boldsymbol{\psi}(\boldsymbol{y}_i, \boldsymbol{W}_i, \boldsymbol{\theta})] = \mathrm{E}[\boldsymbol{\psi}_i(\boldsymbol{\theta})] = \mathbf{0}$, where $\boldsymbol{\psi}$ is a $F \times 1$ vector of functions, $F \geq K+1$. The sample counterpart of these moment conditions will be denoted $\boldsymbol{f}_n(\boldsymbol{\theta}) = 1/n \sum_{i=1}^n \boldsymbol{\psi}_i(\boldsymbol{\theta})$. We thus assume that

A.3 Estimator $\hat{\boldsymbol{\theta}}_n^0$ is $\sqrt{n}$-consistent and asymptotically normal for a fixed T and $n \to \infty$; in particular, $\hat{\boldsymbol{\theta}}_n^0 \xrightarrow{p} \boldsymbol{\theta}^0$ in probability and $\sqrt{n}(\hat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}^0) = O_p(1)$.

The instruments used in this class of the LD estimators can be generically denoted as $\hat{\boldsymbol{z}}_{its} = \boldsymbol{z}_{its} - \boldsymbol{\mathcal{W}}_{its}(\hat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta})$, where explanatory variables $\boldsymbol{\mathcal{W}}_{its} = (\mathbf{0}, \boldsymbol{w}_{i(t-1)}, \cdots, \boldsymbol{w}_{i(t+1-s)})'$ with $\mathbf{0}$ being an appropriately sized matrix of zeros, the initial estimator is assumed to be asymptotically linear in its moment conditions,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}^0) = \boldsymbol{\Lambda} \cdot \sqrt{n} \boldsymbol{f}_n(\boldsymbol{\theta}^0) + o_p(1) = \boldsymbol{\Lambda} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{\psi}_i(\boldsymbol{\theta}^0) + o_p(1), \quad (4.26)$$

and $\boldsymbol{\Lambda}$ is the result of the stochastic expansion of the initial estimator (see for example Arellano, 2003, p. 187).[2]

Next, let $\mathrm{E}[\boldsymbol{\tau}_i(\boldsymbol{\theta})] = \mathbf{0}$ be a general expression for the R moment conditions implied by the method, $R \geq K+1$: after substituting for $\Delta^s \varepsilon_{it}$ from model (4.25), it consists of (4.5) for LD, (4.14) for PD-LD, or (4.17) for MD-LD, respectively, and additionally, of moment conditions implied by $\boldsymbol{x}$'s variables (in general, these depend on whether each x_{itk} is weakly or strictly exogenous or predetermined). Further, the combined vector of moment conditions for the initial and chosen LD-type estimators will be denoted $\boldsymbol{\rho}_i(\boldsymbol{\theta}) = (\boldsymbol{\tau}_i(\boldsymbol{\theta})', \boldsymbol{\psi}_i(\boldsymbol{\theta})')'$. By Assumption A.1, $\boldsymbol{\rho}_1(\boldsymbol{\theta}), \ldots, \boldsymbol{\rho}_n(\boldsymbol{\theta})$ are n independent random vectors. We however have to impose additional assumptions, again taking into account the individual heterogeneity.

A.4 The moment conditions $\boldsymbol{\rho}_i(\boldsymbol{\theta}^0)$ at $\boldsymbol{\theta}^0$ have uniformly bounded finite $(2+\delta)$th moments for some $\delta > 0$. Moreover, $E[\boldsymbol{\rho}_i(\boldsymbol{\theta}^0)] = 0$ and the

[2] Suppose the Arellano-Bond estimator is chosen as preliminary estimator. Let $\boldsymbol{Z}_i^{\mathrm{AB}}$ and $\boldsymbol{A}_n$ denote the corresponding matrix of instruments and the weight matrix, respectively. Then, the matrix $\boldsymbol{\Lambda}$ will be the probability limit of

$$\boldsymbol{\Lambda}_n = -n \left[\sum_{i=1}^n (\boldsymbol{Z}_i^{\mathrm{AB}\prime} \boldsymbol{D}_1 \boldsymbol{W}_i)' \boldsymbol{A}_n \sum_{i=1}^n (\boldsymbol{Z}_i^{\mathrm{AB}\prime} \boldsymbol{D}_1 \boldsymbol{W}_i) \right]^{-1} \sum_{i=1}^n (\boldsymbol{Z}_i^{\mathrm{AB}\prime} \boldsymbol{D}_1 \boldsymbol{W}_i)' \boldsymbol{A}_n,$$

where $\boldsymbol{D}_1$ is the $(T-1) \times T$ first difference-operator matrix.

variance matrix $\boldsymbol{\Sigma} = \lim_{n\to\infty} \sum_{i=1}^{n} \text{var}[\boldsymbol{\rho}_i(\boldsymbol{\theta}^0)]/n$ exists and is a finite positive definite matrix.

A.5 (a) Let $\boldsymbol{\omega}_{tsk} = \lim_{n\to\infty} \sum_{i=1}^{n} \text{E}(\boldsymbol{z}_{its}\Delta^s w_{itk})/n$ exist and be finite for all t, s, and k, and additionally, let $\boldsymbol{\omega}_k = (\boldsymbol{\omega}'_{(S+1)Sk}, \ldots, \boldsymbol{\omega}'_{T(T-1)k})'$ be the kth column of the full-rank matrix $\boldsymbol{\Omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_{K+1})$.
(b) Similarly, let $\boldsymbol{P}_{ts} = \lim_{n\to\infty} \sum_{i=1}^{n} \text{E}(\boldsymbol{\mathcal{W}}_{its}\Delta^s \varepsilon_{it})/n$ exist and be finite for all t and s and let $\boldsymbol{P} = (\boldsymbol{P}'_{(S+1)S}, \ldots, \boldsymbol{P}'_{T(T-1)})'$ have a full rank.
(c) Matrix $\boldsymbol{\Lambda}$ has a full rank.
(d) Finally, $\sum_{i=1}^{n} \text{E}(\boldsymbol{\mathcal{W}}_{its}\Delta^s w_{itk})/n$ is assumed to exist and to be uniformly bounded in $n \in \mathbb{N}$ for all s, t, and k.

A.6 Let $\hat{\boldsymbol{V}}_n$ be a $\dim(\boldsymbol{\tau}) \times \dim(\boldsymbol{\tau})$ inverse weight matrix such that $\hat{\boldsymbol{V}}_n \xrightarrow{p} \boldsymbol{V}$ as $n \to \infty$, where $\boldsymbol{V}$ is a positive definite matrix.

If the standard, but stronger assumption that random variables in Assumption A.1 are independent and identically distributed is used, the above mentioned assumptions would simplify: for example, the moment conditions $\boldsymbol{\rho}_i(\boldsymbol{\theta}^0)$ would have to possess only finite second moments and their variance matrix would be defined simply as $\boldsymbol{\Sigma} = \text{var}[\boldsymbol{\rho}_i(\boldsymbol{\theta}^0)]$.

Under the above stated assumptions, the asymptotic distribution of the feasible LD, MD-LD, and PD-LD estimators can be derived.

Theorem 13. *Suppose that Assumptions A.1–A.6 hold. Then for a fixed T and $n \to \infty$, $\hat{\boldsymbol{\theta}}_n$ is consistent and asymptotically normal:*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0) \xrightarrow{d} \text{N}(\boldsymbol{0}, \boldsymbol{\Xi}), \tag{4.27}$$

where $\boldsymbol{\Xi} = (\boldsymbol{\Omega}'\boldsymbol{V}^{-1}\boldsymbol{\Omega})^{-1}\boldsymbol{\Omega}'\boldsymbol{V}^{-1}\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}'\boldsymbol{V}^{-1}\boldsymbol{\Omega}(\boldsymbol{\Omega}'\boldsymbol{V}^{-1}\boldsymbol{\Omega})^{-1}$ and $\boldsymbol{M} = (\boldsymbol{I}_R, -\boldsymbol{P}\boldsymbol{\Lambda})$.

4.4.2 Estimating the asymptotic variance

According to the standard GMM theory, an optimal choice of the inverse weight matrix $\boldsymbol{V}_n$ is a consistent estimate of the covariance matrix of the moment conditions $\boldsymbol{\Sigma}$. Assuming for simplicity that data are independent and identically distributed across individuals, this covariance matrix can be written as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\tau} & \boldsymbol{\Sigma}_{\tau\psi} \\ \boldsymbol{\Sigma}_{\psi\tau} & \boldsymbol{\Sigma}_{\psi} \end{pmatrix}, \tag{4.28}$$

where $\boldsymbol{\Sigma}_{\tau} = \mathrm{var}[\boldsymbol{\tau}_i(\boldsymbol{\theta}^0)]$, $\boldsymbol{\Sigma}_{\psi} = \mathrm{var}[\boldsymbol{\psi}_i(\boldsymbol{\theta}^0)]$, and $\boldsymbol{\Sigma}_{\tau\psi} = \mathrm{cov}[\boldsymbol{\tau}_i(\boldsymbol{\theta}^0), \boldsymbol{\psi}_i(\boldsymbol{\theta}^0)]$ (recall that $\boldsymbol{\tau}$ and $\boldsymbol{\psi}$ refer to the moment conditions of the (PD-)LD and initial estimators, respectively). Since the instruments are estimated rather than given, Theorem 13 implies that $\boldsymbol{V}_{\mathrm{opt}}$ will be equal to

$$\boldsymbol{V}_{\mathrm{opt}} = \boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}' = \boldsymbol{\Sigma}_{\tau} - \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{\Sigma}_{\psi\tau} - (\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{\Sigma}_{\psi\tau})' + \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{\Sigma}_{\psi}\boldsymbol{\Lambda}'\boldsymbol{P}'. \tag{4.29}$$

Considering the part $\boldsymbol{\Sigma}_{\tau}$, which corresponds to the variance of the moment conditions of the infeasible estimator, note that, because of the complex structure of PD-LD, the covariance matrix $\boldsymbol{\Sigma}_{\tau} = \mathrm{E}(\boldsymbol{Z}_i'\boldsymbol{D}\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'\boldsymbol{D}'\boldsymbol{Z}_i)$ may be singular. In other words, for a sufficiently large T and number of included instruments in PD-LD, some moment conditions are redundant and $\boldsymbol{\Sigma}_{\tau}$ is not invertible.

To overcome this problem in computing $\boldsymbol{V}_{\mathrm{opt}}$, several solutions are available. First, one could try to keep all the moment conditions corresponding to $\boldsymbol{\Sigma}_{\tau}$. This however requires dealing with many linearly dependent moment conditions, which would have to be done as in Carrasco and Florens (2000), for instance. A simple alternative solution – also used in this paper – is to limit the number of instruments in $\boldsymbol{\tau}_i(\boldsymbol{\theta}) = \boldsymbol{Z}_i'\boldsymbol{D}(\boldsymbol{y}_i - \boldsymbol{W}_i\boldsymbol{\theta})$, which equals $\boldsymbol{\tau}_i(\boldsymbol{\theta}^0) = \boldsymbol{Z}_i'\boldsymbol{D}\boldsymbol{\varepsilon}_i$ at $\boldsymbol{\theta}^0$.[3] Denoting $\boldsymbol{\tau}_i^{\dagger}(\boldsymbol{\theta})$ the vector of moment conditions corresponding to selected instruments and $\boldsymbol{Z}_i^{\dagger}$ the corresponding matrix of instruments, the optimal inverse weight matrix will be a consistent estimate

[3] Clearly, there are more ways to do so. To prevent the linear dependence of the PD-LD moment conditions and instruments thereof are selected here in the following way:

$$\mathrm{E}(y_{i(t-s-1)}\Delta^s\varepsilon_{it}) = 0 \qquad (t = s+1,\ldots,T; \quad s = S,\ldots,T-1) \tag{4.30}$$

and for all $t = s+1,\ldots,T$, $s = S,\ldots,T-1$:

$$\begin{cases} \mathrm{E}(u_{i(t-1)}\Delta^s\varepsilon_{it}) = 0 & \text{if } s > S; \\ \mathrm{E}\left[\begin{pmatrix} u_{i(t-1)} \\ \vdots \\ u_{i(t-s+1)} \end{pmatrix}\Delta^s\varepsilon_{it}\right] = \mathbf{0} & \text{if } s = S. \end{cases} \tag{4.31}$$

.

$\hat{\boldsymbol{V}}_{n\Upsilon}$ of (4.29), where

$$\hat{\boldsymbol{\Sigma}}_{n\boldsymbol{\Upsilon}} = \sum_{i=1}^{n} \hat{\boldsymbol{Z}}_i^{\dagger\prime} \hat{\boldsymbol{\Upsilon}}_i \hat{\boldsymbol{Z}}_i^{\dagger} \tag{4.32}$$

with $\hat{\boldsymbol{\Upsilon}}_i = \boldsymbol{D}\hat{\boldsymbol{\varepsilon}}_i\hat{\boldsymbol{\varepsilon}}_i'\boldsymbol{D}'$ and $\hat{\boldsymbol{Z}}_i^{\dagger}$ being computed by using a preliminary consistent estimator ($\boldsymbol{\Lambda}$, $\boldsymbol{P}$, and other terms in (4.29) can be estimated by the respective sample averages).

For several reasons, we do not pay more attention to the estimation of the optimal weights $\boldsymbol{V}_{\text{opt}}$. It is well known that a part of the bias of GMM estimators stems from a poorly estimated weight matrix (Newey and Smith, 2004). For either small values of n or large number of instruments (which depends on T when all instruments are included), weights in $\hat{\boldsymbol{V}}_{n\Upsilon}$ may be highly imprecise. A simple alternative to (4.32) is to employ the weights of the standard two-stage least squares and use instead

$$\hat{\boldsymbol{V}}_{nI} = \sum_{i=1}^{n} \hat{\boldsymbol{Z}}_i' \boldsymbol{I} \hat{\boldsymbol{Z}}_i. \tag{4.33}$$

There are a couple of advantages of this weighting matrix $\hat{\boldsymbol{V}}_{nI}$: (i) it can be computed directly based on the initial estimate, (ii) it does not impose constraints on the number of instruments (the full proposed matrix $\boldsymbol{Z}_i$ can be used), and finally, (iii) finite sample results for weighting matrix (4.33) are rather close to or even better than the ones for weighting matrix (4.32), especially as the sample size increases. See Section 4.5 for more details.

4.5 Monte Carlo simulation

4.5.1 Design

In this section, the finite sample performance of the proposed estimators is evaluated by Monte Carlo simulations. The data-generating proces for y_{it} follows model (4.1) with $\alpha = 0.1, 0.5, 0.9$, $n = 25, 50, 100, 400, 1600, 3200$, $T = 6, 12, 24$, $\eta_i \sim \mathrm{N}(0, \sigma_\eta^2)$, and $\varepsilon_{it} \sim \mathrm{N}(0, 1)$. In order to measure the sensitivity of the estimators to the stationarity assumption, the initial observations at time $t = 0$ are generated by

$$y_{i0} \sim \mathrm{N}\left(\frac{\eta_i}{1-\alpha_J}, \frac{\sigma_\varepsilon^2}{1-\alpha^2}\right), \tag{4.34}$$

which leads to mean-stationary series y_{it} if $\alpha_J = \alpha$ and to non-stationary sequences if $\alpha_J \neq \alpha$. Each model is evaluated using 1000 replications.

Results are reported for the LD estimator and for the proposed estimators MD-LD, PD-LD, and PD-LD1, where the last one denotes the iterated PD-LD estimator based on PD-LD used as the preliminary estimator. The Arellano and Bond (AB, 1991) two-step GMM estimator[4] and the system Blundell and Bond (BB, 1998) estimator[5] are also reported, serving as reference estimators as well as preliminary estimators for LD, MD-LD, and PD-LD. All methods are compared by means of the root mean squared errors (RMSE) unless stated otherwise.

4.5.2 Weight matrix

Before presenting a full comparison of estimators, we will briefly revisit the choice of the GMM weight matrix. As mentioned in Section 4.4.2, the finite sample performance of a GMM estimator can be heavily affected by the choice of the weight matrix. The difference between using weights (4.32) and (4.33) is documented in Table 4.2 for various models with $\sigma_\eta^2 = 1$. Let PD-LD-I and PD-LD1-I denote the PD-LD estimators when the inverse weight matrix (4.33) is used and let PD-LD-$\hat{\Upsilon}$ and PD-LD1-$\hat{\Upsilon}$ denote PD-LD when weights (4.32) are in use.

As shown in Table 4.2, PD-LD-$\hat{\Upsilon}$ seems to perform only slightly better than PD-LD-I and only for small values of n (the main exception is the case of $n = 25$, $T = 6$, and $\alpha = 0.9$). More specifically, PD-LD-$\hat{\Upsilon}$ can perform slightly better than PD-LD-I if the initial estimator is reliable, but PD-LD-$\hat{\Upsilon}$ can perform much worse than PD-LD-I if the initial estimator is imprecise. Consequently, it seems that using weights (4.33) is a more robust strategy, which – in the cases when it is worse than PD-LD-$\hat{\Upsilon}$ – matches the optimally weighted alternative once the sample size is sufficiently large. We therefore recommend and use in further simulations the PD-LD estimator based on the weighting matrix $\hat{\boldsymbol{V}}_{nI}$ defined in (4.33).

[4]The (optimal) inverse weight matrix is $\sum_i \boldsymbol{Z}_i^{\mathrm{AB}\prime} \boldsymbol{H} \boldsymbol{Z}_i^{\mathrm{AB}}$, where $\boldsymbol{Z}_i^{\mathrm{AB}}$ is the matrix of instruments and $\boldsymbol{H}$ is a $(T-1)\times(T-1)$ tridiagonal matrix with 2 on the main diagonal, -1 on the first two sub-diagonals, and zeros elsewhere (see Arellano and Bond, 1991, p. 279).

[5]The inverse weight matrix is $\sum_i \boldsymbol{Z}_i^{\mathrm{BB}\prime} \boldsymbol{G} \boldsymbol{Z}_i^{\mathrm{BB}}$, where $\boldsymbol{Z}_i^{\mathrm{BB}}$ is the matrix of instruments and $\boldsymbol{G}$ is a partitioned matrix, $\boldsymbol{G} = \mathrm{diag}(\boldsymbol{H}, \boldsymbol{I})$, where $\boldsymbol{H}$ is as in Arellano-Bond and $\boldsymbol{I}$ is the identity matrix (see Kiviet, 2007, Eq. (38)).

Table 4.2: The root mean squared errors of the PD-LD estimators using the two-stage least-squares weighting matrix and the asymptotically optimal weighting matrix. The three sections of the table represent results for $\alpha = 0.1, 0.5$, and 0.9.

n	25	100	400	1600	3200	25	100	25	100
T			6			12		24	
				$\alpha = 0.1$					
AB*	0.141	0.071	0.035	0.017	0.012	0.092	0.041	0.063	0.028
PD-LD-I	0.138	0.067	0.034	0.016	0.012	0.095	0.047	0.076	0.039
PD-LD-$\hat{\Upsilon}$	0.135	0.063	0.032	0.016	0.011	0.095	0.042	0.076	0.039
PD-LD1-I	0.141	0.067	0.034	0.017	0.012	0.095	0.047	0.076	0.039
PD-LD1-$\hat{\Upsilon}$	0.157	0.076	0.038	0.018	0.013	0.095	0.053	0.076	0.039
BB*	0.127	0.067	0.035	0.017	0.013	0.100	0.046	0.157	0.052
PD-LD-I	0.138	0.065	0.033	0.017	0.012	0.093	0.048	0.075	0.038
PD-LD-$\hat{\Upsilon}$	0.127	0.062	0.031	0.016	0.012	0.093	0.046	0.075	0.038
PD-LD1-I	0.141	0.066	0.033	0.017	0.012	0.093	0.048	0.075	0.038
PD-LD1-$\hat{\Upsilon}$	0.153	0.074	0.037	0.018	0.013	0.093	0.057	0.075	0.038
				$\alpha = 0.5$					
AB*	0.232	0.108	0.051	0.025	0.018	0.133	0.055	0.080	0.033
PD-LD-I	0.136	0.072	0.035	0.018	0.013	0.090	0.045	0.069	0.037
PD-LD-$\hat{\Upsilon}$	0.187	0.076	0.034	0.017	0.012	0.090	0.047	0.069	0.037
PD-LD1-I	0.155	0.082	0.040	0.020	0.014	0.092	0.046	0.069	0.037
PD-LD1-$\hat{\Upsilon}$	0.183	0.084	0.040	0.020	0.014	0.092	0.053	0.069	0.037
BB*	0.139	0.081	0.044	0.021	0.016	0.118	0.058	0.186	0.074
PD-LD-I	0.135	0.069	0.034	0.017	0.012	0.087	0.045	0.070	0.035
PD-LD-$\hat{\Upsilon}$	0.131	0.067	0.033	0.017	0.012	0.087	0.049	0.070	0.035
PD-LD1-I	0.157	0.080	0.039	0.019	0.014	0.088	0.046	0.070	0.035
PD-LD1-$\hat{\Upsilon}$	0.152	0.078	0.039	0.020	0.014	0.088	0.054	0.070	0.035
				$\alpha = 0.9$					
AB*	0.570	0.444	0.241	0.102	0.069	0.292	0.202	0.146	0.089
PD-LD-I	0.202	0.160	0.120	0.072	0.050	0.093	0.064	0.049	0.026
PD-LD-$\hat{\Upsilon}$	0.439	0.268	0.128	0.065	0.046	0.093	0.155	0.049	0.026
PD-LD1-I	0.186	0.127	0.097	0.072	0.044	0.089	0.050	0.049	0.026
PD-LD1-$\hat{\Upsilon}$	0.389	0.205	0.099	0.063	0.043	0.089	0.129	0.049	0.026
BB*	0.082	0.073	0.051	0.030	0.021	0.054	0.048	0.043	0.029
PD-LD-I	0.124	0.085	0.052	0.028	0.019	0.067	0.042	0.047	0.023
PD-LD-$\hat{\Upsilon}$	0.109	0.081	0.053	0.029	0.020	0.067	0.043	0.047	0.023
PD-LD1-I	0.207	0.129	0.067	0.033	0.022	0.091	0.052	0.050	0.027
PD-LD1-$\hat{\Upsilon}$	0.161	0.111	0.063	0.032	0.021	0.091	0.046	0.050	0.027

Note: The symbol '*' denotes the preliminary estimator for PD-LD.

Table 4.3: The root mean squared errors of all estimator for different sample sizes using $\sigma_\eta^2 = 1$. The three sections of the table represent results for $\alpha = 0.1, 0.5$, and 0.9.

n	25			50			100		
T	6	12	24	6	12	24	6	12	24
				$\alpha = 0.1$					
AB*	0.143	0.089	0.064	0.097	0.058	0.041	0.070	0.041	0.027
LD	0.206	0.209	0.206	0.134	0.148	0.141	0.099	0.103	0.100
MD-LD	0.145	0.120	0.110	0.100	0.081	0.078	0.072	0.059	0.052
PD-LD	0.136	0.094	0.074	0.093	0.067	0.053	0.070	0.047	0.037
PD-LD1	0.139	0.094	0.074	0.094	0.067	0.053	0.071	0.047	0.037
BB*	0.129	0.102	0.157	0.095	0.067	0.091	0.068	0.046	0.050
LD	0.206	0.209	0.206	0.134	0.148	0.141	0.099	0.103	0.100
MD-LD	0.145	0.119	0.110	0.099	0.081	0.078	0.071	0.059	0.052
PD-LD	0.136	0.094	0.074	0.093	0.067	0.053	0.070	0.047	0.037
PD-LD1	0.139	0.094	0.074	0.094	0.067	0.053	0.071	0.047	0.037
				$\alpha = 0.5$					
AB*	0.231	0.129	0.083	0.152	0.083	0.053	0.107	0.053	0.033
LD	0.188	0.184	0.174	0.124	0.127	0.121	0.091	0.087	0.086
MD-LD	0.155	0.112	0.099	0.112	0.076	0.069	0.079	0.054	0.046
PD-LD	0.138	0.088	0.070	0.098	0.064	0.049	0.072	0.043	0.035
PD-LD1	0.160	0.090	0.070	0.109	0.065	0.049	0.083	0.044	0.035
BB*	0.148	0.119	0.184	0.113	0.084	0.127	0.082	0.057	0.074
LD	0.191	0.184	0.174	0.123	0.127	0.121	0.091	0.087	0.086
MD-LD	0.154	0.112	0.099	0.107	0.076	0.069	0.076	0.054	0.046
PD-LD	0.141	0.088	0.070	0.095	0.064	0.049	0.072	0.043	0.035
PD-LD1	0.165	0.090	0.070	0.109	0.065	0.049	0.084	0.044	0.035
				$\alpha = 0.9$					
AB*	0.579	0.296	0.146	0.516	0.256	0.122	0.447	0.201	0.089
LD	0.197	0.119	0.096	0.169	0.088	0.066	0.146	0.066	0.045
MD-LD	0.233	0.125	0.073	0.207	0.099	0.056	0.182	0.082	0.040
PD-LD	0.201	0.093	0.047	0.180	0.076	0.035	0.158	0.063	0.026
PD-LD1	0.189	0.089	0.048	0.152	0.068	0.037	0.124	0.051	0.026
BB*	0.086	0.053	0.041	0.079	0.052	0.036	0.070	0.047	0.029
LD	0.171	0.114	0.096	0.120	0.084	0.065	0.095	0.059	0.045
MD-LD	0.142	0.085	0.065	0.108	0.065	0.047	0.085	0.053	0.034
PD-LD	0.130	0.064	0.045	0.094	0.052	0.032	0.080	0.041	0.022
PD-LD1	0.212	0.088	0.048	0.154	0.073	0.037	0.121	0.052	0.026

Note: The symbol '*' denotes the preliminary estimator for LD, MD-LD and PD-LD.

Table 4.4: The root mean squared errors of all estimator for low and high values of the variance of the individual effects, $\sigma_\eta^2 = 0.25, 1, 4$. The three sections of the table represent results for $\alpha = 0.1, 0.5$, and 0.9.

(n, T)	$(100, 6)$			$(50, 12)$			$(25, 24)$		
$\sigma_\eta^2/\sigma_\varepsilon^2$	1/4	1	4	1/4	1	4	1/4	1	4
				$\alpha = 0.1$					
AB*	0.062	0.070	0.076	0.053	0.062	0.062	0.063	0.062	0.064
LD	0.097	0.104	0.101	0.145	0.148	0.144	0.213	0.208	0.211
MD-LD	0.070	0.071	0.069	0.082	0.084	0.082	0.106	0.109	0.107
PD-LD	0.066	0.069	0.068	0.066	0.066	0.066	0.070	0.075	0.074
PD-LD1	0.066	0.070	0.069	0.066	0.066	0.066	0.070	0.075	0.074
BB*	0.060	0.068	0.081	0.074	0.069	0.071	0.176	0.159	0.091
LD	0.100	0.102	0.100	0.145	0.141	0.138	0.201	0.200	0.203
MD-LD	0.069	0.069	0.070	0.084	0.080	0.083	0.104	0.107	0.108
PD-LD	0.067	0.067	0.065	0.065	0.065	0.065	0.073	0.075	0.075
PD-LD1	0.068	0.068	0.066	0.065	0.065	0.065	0.073	0.075	0.075
				$\alpha = 0.5$					
AB*	0.086	0.108	0.122	0.071	0.087	0.092	0.080	0.082	0.085
LD	0.089	0.093	0.091	0.127	0.128	0.123	0.183	0.175	0.178
MD-LD	0.079	0.079	0.077	0.079	0.078	0.078	0.094	0.095	0.096
PD-LD	0.069	0.070	0.069	0.063	0.061	0.062	0.067	0.071	0.070
PD-LD1	0.078	0.078	0.080	0.064	0.062	0.064	0.067	0.071	0.070
BB*	0.076	0.083	0.118	0.107	0.087	0.104	0.236	0.187	0.083
LD	0.094	0.092	0.098	0.128	0.123	0.120	0.175	0.176	0.175
MD-LD	0.077	0.077	0.089	0.080	0.075	0.081	0.095	0.098	0.096
PD-LD	0.070	0.069	0.077	0.060	0.061	0.061	0.069	0.071	0.071
PD-LD1	0.081	0.080	0.083	0.062	0.062	0.062	0.069	0.071	0.071
				$\alpha = 0.9$					
AB*	0.345	0.470	0.494	0.221	0.257	0.270	0.139	0.147	0.148
LD	0.139	0.151	0.152	0.088	0.089	0.086	0.092	0.094	0.098
MD-LD	0.165	0.186	0.192	0.098	0.101	0.103	0.072	0.074	0.075
PD-LD	0.146	0.163	0.166	0.076	0.076	0.076	0.045	0.048	0.048
PD-LD1	0.130	0.146	0.129	0.068	0.067	0.067	0.046	0.049	0.049
BB*	0.076	0.071	0.088	0.086	0.051	0.081	0.154	0.044	0.066
LD	0.095	0.097	0.114	0.083	0.081	0.093	0.100	0.093	0.092
MD-LD	0.089	0.089	0.104	0.074	0.064	0.083	0.076	0.066	0.062
PD-LD	0.082	0.082	0.098	0.057	0.048	0.064	0.048	0.047	0.044
PD-LD1	0.118	0.127	0.152	0.065	0.066	0.078	0.049	0.050	0.050

Note: The symbol '*' denotes the preliminary estimator for LD, MD-LD and PD-LD.

Table 4.5: The biases and root mean squared errors of all estimator under the non-stationarity of the initial condition; $\sigma_\eta^2 = 1$ and $\alpha_J = 0.3$ are used. The three sections of the table represent results for $\alpha = 0.1, 0.5$, and 0.9.

(n, T)	$(100, 6)$		$(50, 12)$		$(25, 24)$	
	RMSE	Bias	RMSE	Bias	RMSE	Bias
			$\alpha = 0.1$			
AB*	0.064	-0.019	0.059	-0.027	0.062	-0.044
LD	0.096	-0.002	0.142	0.001	0.198	0.010
MD-LD	0.072	-0.007	0.083	-0.004	0.103	0.002
PD-LD	0.065	-0.003	0.066	-0.001	0.075	0.003
PD-LD1	0.066	-0.003	0.066	-0.001	0.075	0.003
BB*	0.065	-0.018	0.070	-0.042	0.158	-0.150
LD	0.096	0.006	0.147	0.001	0.205	-0.013
MD-LD	0.065	-0.002	0.082	-0.004	0.103	-0.005
PD-LD	0.065	0.001	0.068	-0.003	0.072	-0.002
PD-LD1	0.065	0.002	0.068	-0.003	0.072	-0.002
			$\alpha = 0.5$			
AB*	0.139	-0.071	0.094	-0.067	0.088	-0.076
LD	0.087	-0.008	0.117	-0.004	0.170	0.002
MD-LD	0.077	-0.019	0.073	-0.013	0.094	-0.005
PD-LD	0.068	-0.013	0.059	-0.008	0.070	-0.006
PD-LD1	0.076	-0.006	0.061	-0.008	0.070	-0.006
BB*	0.125	0.094	0.070	-0.015	0.174	-0.164
LD	0.090	0.024	0.115	-0.006	0.170	-0.004
MD-LD	0.085	0.040	0.074	-0.008	0.097	-0.007
PD-LD	0.076	0.030	0.058	-0.005	0.071	-0.005
PD-LD1	0.077	0.007	0.059	-0.005	0.071	-0.005
			$\alpha = 0.9$			
AB*	0.072	-0.029	0.052	-0.034	0.056	-0.048
LD	0.044	-0.007	0.032	-0.004	0.035	-0.000
MD-LD	0.042	-0.007	0.030	-0.006	0.031	-0.003
PD-LD	0.041	-0.008	0.024	-0.006	0.020	-0.003
PD-LD1	0.039	-0.004	0.023	-0.004	0.020	-0.003
BB*	0.208	0.207	0.146	0.145	0.072	0.063
LD	0.081	0.069	0.041	0.025	0.035	-0.001
MD-LD	0.086	0.075	0.043	0.030	0.031	0.001
PD-LD	0.082	0.073	0.038	0.030	0.019	-0.002
PD-LD1	0.046	0.018	0.024	0.001	0.020	-0.003

Note: The symbol '*' denotes the preliminary estimator for LD, MD-LD and PD-LD. The initial observations are generated by $y_{i0} \sim \mathrm{N}(\frac{\eta_i}{1-0.3}, \frac{\sigma_\varepsilon^2}{1-\alpha^2})$.

4.5.3 Simulation results

First, an overview of the behaviour of all estimators is given for many different sample sizes and $\sigma_\eta^2/\sigma_\varepsilon^2 = 1$, see Table 4.3. By taking only the difference between the last and the first observation per individual, the LD estimator yields almost no benefit from a larger number T of time periods, unless $\alpha = 0.9$ (then there are more available informative instruments as T increases). In particular, LD performs poorly when α is close to zero. These weaknesses are amended by the proposed estimators. Among these, PD-LD has an overall good performance for all combinations of n and T: (i) it always performs better than LD and MD-LD; (ii) it exhibits smaller RMSEs than AB for $\alpha \geq 0.5$ and is rather close to AB for $\alpha = 0.1$; and (iii) it outperforms BB for $\alpha \leq 0.5$ and – if BB is the initial estimator – PD-LD has similar or smaller RMSE compared to BB for $\alpha = 0.9$ except for the smallest sample size $n = 25$ and $T = 6$. Finally, it is interesting to note that the precision of the PD-LD estimates does not depend much on the initial estimator except for $\alpha = 0.9$, where AB gets very imprecise and substantially biased.

Further, the estimators in the LD class also do not seem to be affected by different ratios of $\sigma_\eta^2/\sigma_\varepsilon^2$. This is documented in Table 4.4. In the performed experiments, the AB estimator is not substantially influenced by variations in the ratio $\sigma_\eta^2/\sigma_\varepsilon^2$ either. On the contrary, the BB estimator is the most sensitive, in particular when T is large.

Finally, we examine the sensitivity of the estimators to misspecification of the initial condition assumption; Table 4.5 summarizes now both the RMSE and biases for all estimates. The initial observations y_{i0} are defined as in (4.34) and $\alpha_J = 0.3$ for all $\alpha \in \{0.1, 0.5, 0.9\}$. It is well known that the BB estimator loses its predominant source of information when y_{it} is mean-nonstationary (see Hahn, 1999). On the contrary, all estimators in the LD class are not substantially affected by different assumptions about y_{i0}. In particular, the biases of LD estimators are almost zero if AB is used as the initial estimator. (Note that the AB estimator actually benefits from mean-nonstationarity, especially when α is close to one, as documented in Hayakawa (2009).) In the other case of the initial BB estimator, LD and PD-LD substantially reduce the bias of the initial estimator, and surprisingly, PD-LD1 even manages to eliminate the bias almost completely (i.e., despite a sizeable upward bias of BB for $\alpha = 0.9$). Finally, note that PD-LD and PD-LD1 exhibit the smallest RMSE of all estimators if $\alpha \geq 0.5$.

Altogether, the PD-LD estimator performs equally well or better than existing methods in the majority of simulated models. The reported experiments show that these results are not overly sensitive to the values of the autoregressive parameter, to the variance of errors, or to the specification of initial observations.

4.6 Conclusion

To our knowledge, the idea of applying multiple pairwise differences to dynamic linear panel data models is new. This data transformation is presented and applied here to the long-difference estimator of Hahn et al. (2007) to improve its behavior for data with many time periods and for the values of the autoregressive coefficient far from one. We derive the finite-sample bias of the method and the asymptotic distribution of the proposed estimators. Our results indicate that the PD-LD estimator has a smaller variance than the original LD estimator, while preserving its very small bias. In finite samples, simulation results confirm that the proposed pairwise-difference transformation improves the LD estimator in all simulation settings, and in particular, when the time span increases or when α is small. Compared to the existing IV/GMM type of estimator, PD-LD seems to be very competitive without imposing additional restrictive assumptions.

4.A Appendix

4.A.1 Finite sample bias

Let us first state and prove the following lemma, which will be used for evaluating of the bias expressions.

Lemma 3. *Let $J \in \mathbb{N}$ and $|\gamma| < 1$. Then it holds for $1 \le K < L, K \in \mathbb{N}, L \in \mathbb{N}$, that*

$$\sum_{j=K}^{L} j\gamma^j = \frac{\gamma^K - \gamma^{L+1}}{(1-\gamma)^2} - \frac{L\gamma^{L+1} - (K-1)\gamma^K}{1-\gamma}, \tag{4.35}$$

$$\sum_{j=K}^{L} j^2\gamma^j = 2\frac{\gamma^K - \gamma^{L+1}}{(1-\gamma)^3} - \frac{(2L-1)\gamma^{L+1} - (2K-3)\gamma^K}{(1-\gamma)^2} - \frac{L^2\gamma^{L+1} - (K-1)^2\gamma^K}{1-\gamma}. \tag{4.36}$$

Proof. The proof follows directly from $(1-\gamma)\sum_{j=0}^{J}\gamma = 1-\gamma^{J+1}$:

$$\sum_{j=0}^{J} j\gamma^j = \sum_{j=1}^{J}\sum_{l=j}^{J}\gamma^l = \sum_{j=1}^{J}\gamma^j\frac{1-\gamma^{J-j+1}}{1-\gamma} = \frac{1}{1-\gamma}\left(\sum_{j=0}^{J}\gamma^j - 1 - J\gamma^{J+1}\right)$$
$$= \frac{1-\gamma^{J+1}}{(1-\gamma)^2} - \frac{J\gamma^{J+1}+1}{1-\gamma},$$

and using the above result,

$$\sum_{j=0}^{J} j^2\gamma^j = \sum_{j=1}^{J}[\sum_{l=1}^{j}(2l-1)]\gamma^l = \sum_{j=1}^{J}(2j-1)\sum_{l=j}^{J}\gamma^l = \frac{1}{1-\gamma}\left(\sum_{j=1}^{J}(2j-1)\gamma^j - J^2\gamma^{J+1}\right)$$
$$= 2\frac{1-\gamma^{J+1}}{(1-\gamma)^3} - \frac{(2J-1)\gamma^{J+1}+\gamma+1}{(1-\gamma)^2} - \frac{J^2\gamma^{J+1}}{1-\gamma}.$$

Writing now sums $\sum_{j=K}^{L} a_j$ as $\sum_{j=0}^{L} a_j - \sum_{j=0}^{K-1} a_j$ implies the results of the lemma. $\square$

Proof of Theorem 12. In this proof we follow Bun and Kiviet (2006, Appendix A). Results are fully derived for the infeasible PD-LD estimator only. For LD and MD-LD, the proof develops identically except for the final evaluation of the biases as functions of the autoregressive parameter α. We thus proceed with the proof for PD-LD and only the final evaluation is done for each estimator separately.

The estimation error of the unfeasible PD-LD estimator in (4.16) (obtained after substituting from the model equations (4.15)) is given by

$$\hat{\alpha}_{nT} - \alpha = \frac{\boldsymbol{y}_{-1}^{*\prime}\boldsymbol{Z}\boldsymbol{V}_n^{-1}\boldsymbol{Z}'\boldsymbol{\varepsilon}^*}{\boldsymbol{y}_{-1}^{*\prime}\boldsymbol{Z}\boldsymbol{V}_n^{-1}\boldsymbol{Z}'\boldsymbol{y}_{-1}^*} = \frac{\boldsymbol{g}_{nT^*}'\boldsymbol{\varepsilon}^*}{q_{nT^*}}, \tag{4.37}$$

where $\boldsymbol{\varepsilon}^* = ([\boldsymbol{D}\boldsymbol{\varepsilon}_1]', \ldots, [\boldsymbol{D}\boldsymbol{\varepsilon}_n]')'$ and $\boldsymbol{g}_{nT^*} = \boldsymbol{Z}\boldsymbol{V}_n^{-1}\boldsymbol{Z}'\boldsymbol{y}_{-1}^*$. Suppose that $\mathrm{E}(\boldsymbol{g}_{nT^*}'\boldsymbol{\varepsilon}^*) = O(N^*)$, where N^* is some function of n and/or T^* to be derived yet. Assuming that either or both n and T^* can get large, Bun and Kiviet (2006, Eq. (31)–(33)) showed that the first-order bias approximation of $\hat{\alpha}_{nT}$ is given by

$$\mathrm{E}(\hat{\alpha}_{nT}-\alpha) = \frac{\mathrm{E}(\boldsymbol{g}_{nT^*}'\boldsymbol{\varepsilon}^*)}{\bar{q}} + O(N^*(nT^*)^{-3/2}) = B + O(N^*(nT^*)^{-3/2}), \tag{4.38}$$

where $B = \mathrm{E}(\boldsymbol{g}_{nT^*}'\boldsymbol{\varepsilon}^*)/\bar{q} = O(N^*(nT^*)^{-1})$ is the leading term of the bias. Note that the term $(nT^*)^{-1}$ in the previous expressions follows from Assumption B.3 as $q_{nT^*}/(nT^*) \to \bar{q} > 0$ in probability and $q_{nT^*} = O((nT^*)^{-1})$

for $nT^* \to \infty$.

Next, let us derive N^*. First, we can rewrite (4.37) in a more convenient form. Let $\boldsymbol{G}$ be an $nT^* \times nT^*$ permutation matrix which changes the order of the rows of $\boldsymbol{Z}$, $\boldsymbol{y}^*_{-1}$, $\boldsymbol{y}^*$ such that observations are organized by individuals first ($i = 1, \ldots, n$), then by pairwise differences ($s = S, \ldots, T-1$), and last by time periods ($t = s+1, \ldots T$). As $\boldsymbol{Z}_i$ is block diagonal, $\boldsymbol{G}'\boldsymbol{Z}$ will be block diagonal as well with blocks $\boldsymbol{Z}_{(S+1)S}, \ldots, \boldsymbol{Z}_{T(T-1)}$, where $\boldsymbol{Z}_{ts} = ((\boldsymbol{z}_{its})_{i=1}^n)'$ is $n \times m_{ts}$ and m_{ts} denotes the number of instruments. The inverse weight matrix used here is $\boldsymbol{V}_n = \boldsymbol{Z}'\boldsymbol{Z}$ and is thus also block diagonal. Given that

$$(\boldsymbol{Z}'\boldsymbol{Z})^{-1} = (\boldsymbol{Z}'\boldsymbol{G}'\boldsymbol{G}\boldsymbol{Z})^{-1} = \operatorname{diag}\left((\boldsymbol{Z}'_{(S+1)S}\boldsymbol{Z}_{(S+1)S})^{-1}, \ldots, (\boldsymbol{Z}'_{T(T-1)}\boldsymbol{Z}_{T(T-1)})^{-1}\right), \tag{4.39}$$

we can rewrite (4.37) as

$$\hat{\alpha}_{nT} - \alpha = \frac{\boldsymbol{g}'_{nT^*}\boldsymbol{\varepsilon}^*}{q_{nT^*}} = \frac{\sum_{s=S}^{T-1}\sum_{t=s+1}^{T} \boldsymbol{y}^{*\prime}_{(t-1)s}\boldsymbol{Z}_{ts}(\boldsymbol{Z}'_{ts}\boldsymbol{Z}_{ts})^{-1}\boldsymbol{Z}'_{ts}\boldsymbol{\varepsilon}^*_{ts}}{\sum_{s=S}^{T-1}\sum_{t=s+1}^{T} \boldsymbol{y}^{*\prime}_{(t-1)s}\boldsymbol{Z}_{ts}(\boldsymbol{Z}'_{ts}\boldsymbol{Z}_{ts})^{-1}\boldsymbol{Z}'_{ts}\boldsymbol{y}^*_{(t-1)s}}, \tag{4.40}$$

where $\boldsymbol{y}^*_{(t-1)s} = (y^*_{1(t-1)s}, \ldots, y^*_{n(t-1)s})'$ and $\boldsymbol{\varepsilon}^*_{ts} = (\varepsilon^*_{1ts}, \ldots, \varepsilon^*_{nts})'$ using $y^*_{i(t-1)s} = \Delta^s y_{i(t-1)} = y_{i(t-1)} - y_{i(t-1-s)}$ and $\varepsilon^*_{its} = \Delta^s \varepsilon_{it} = \varepsilon_{it} - \varepsilon_{i(t-s)}$. We now have to analyze the expectation of the nominator of (4.40),

$$\mathrm{E}(\boldsymbol{g}'_{nT^*}\boldsymbol{\varepsilon}^*) = \mathrm{E}\left(\sum_{s=S}^{T-1}\sum_{t=s+1}^{T} \boldsymbol{y}^{*\prime}_{(t-1)s}\boldsymbol{M}_{ts}\boldsymbol{\varepsilon}^*_{ts}\right) = \sum_{s=S}^{T-1}\sum_{t=s+1}^{T} \mathrm{E}(\boldsymbol{y}^{*\prime}_{(t-1)s}\boldsymbol{M}_{ts}\boldsymbol{\varepsilon}^*_{ts}), \tag{4.41}$$

where $\boldsymbol{M}_{ts} = \boldsymbol{Z}_{ts}(\boldsymbol{Z}'_{ts}\boldsymbol{Z}_{ts})^{-1}\boldsymbol{Z}'_{ts}$, with $\operatorname{tr}(\boldsymbol{M}_{ts}) = \operatorname{tr}(\boldsymbol{I}_{m_{ts}}) = m_{ts}$. Next,

$$\begin{aligned}\mathrm{E}(\boldsymbol{y}^{*\prime}_{(t-1)s}\boldsymbol{M}_{ts}\boldsymbol{\varepsilon}^*_{ts}) &= \mathrm{E}\left[\operatorname{tr}(\boldsymbol{y}^{*\prime}_{(t-1)s}\boldsymbol{M}_{ts}\boldsymbol{\varepsilon}^*_{ts})\right] = \mathrm{E}\left[\operatorname{tr}(\boldsymbol{M}_{ts}\boldsymbol{\varepsilon}^*_{ts}\boldsymbol{y}^{*\prime}_{(t-1)s})\right] \\ &= \mathrm{E}\left\{\mathrm{E}\left[\operatorname{tr}(\boldsymbol{M}_{ts}\boldsymbol{\varepsilon}^*_{ts}\boldsymbol{y}^{*\prime}_{(t-1)s})\middle|\mathcal{I}_{t-1}\right]\right\},\end{aligned} \tag{4.42}$$

where $\mathrm{E}(\cdot|\mathcal{I}_{t-1})$ denotes the expectation conditional on the information known up to $t-1$. Note that $\boldsymbol{Z}_{ts}$ and thus $\boldsymbol{M}_{ts}$ contain only relevant stochastic elements that have been observed prior to t. Hence

$$\begin{aligned}
\mathrm{E}\left\{\mathrm{E}\left[\operatorname{tr}(\boldsymbol{M}_{ts}\boldsymbol{\varepsilon}^*_{ts}\boldsymbol{y}^{*\prime}_{(t-1)s})\middle|\,\mathcal{I}_{t-1}\right]\right\} &= \mathrm{E}\left\{\operatorname{tr}\left[\boldsymbol{M}_{ts}\,\mathrm{E}\left(\boldsymbol{\varepsilon}^*_{ts}\boldsymbol{y}^{*\prime}_{(t-1)s}\middle|\,\mathcal{I}_{t-1}\right)\right]\right\} \\
&= \mathrm{E}\left\{\operatorname{tr}\left[\boldsymbol{M}_{ts}\boldsymbol{I}_n\,\mathrm{E}\left(\varepsilon^*_{its}y^*_{i(t-1)s}\middle|\,\mathcal{I}_{t-1}\right)\right]\right\} \\
&= \mathrm{E}\left[\operatorname{tr}\left(\boldsymbol{M}_{ts}\right)\mathrm{E}\left(\varepsilon^*_{its}y^*_{i(t-1)s}\middle|\,\mathcal{I}_{t-1}\right)\right] \\
&= m_{ts}\,\mathrm{E}\left[\mathrm{E}\left(\varepsilon^*_{its}y^*_{i(t-1)s}\middle|\,\mathcal{I}_{t-1}\right)\right],
\end{aligned} \tag{4.43}$$

provided that the conditional expections are independent of index i. Under Assumptions B.1–B.3, this however follows from the definition of the transformed variables $y^*_{i(t-1)s}$ and ε^*_{its}:

$$\begin{aligned}
\mathrm{E}\left[\mathrm{E}\left(\varepsilon^*_{its}y^*_{i(t-1)s}\middle|\,\mathcal{I}_{t-1}\right)\right] &= \mathrm{E}\left[\mathrm{E}\left(\Delta^s\varepsilon_{it}\Delta^s y_{i(t-1)}\middle|\,\mathcal{I}_{t-1}\right)\right] = -\,\mathrm{E}\left(\varepsilon_{i(t-s)}y_{i(t-1)}\right) \\
&= -\,\mathrm{E}\left\{\varepsilon_{i(t-s)}\left[\left(\sum_{k=0}^{t-2}\alpha^k\right)\eta_i + \alpha^{t-1}y_{i0} + \sum_{k=0}^{t-2}\alpha^k\varepsilon_{i(t-1-k)}\right]\right\} \\
&= -\alpha^{s-1}\sigma^2_\varepsilon,
\end{aligned} \tag{4.44}$$

where $\sigma^2_\varepsilon = \operatorname{var}(\varepsilon_{it})$ for all i and t. This implies for equation (4.41) that

$$\sum_{s=S}^{T-1}\sum_{t=s+1}^{T}\mathrm{E}(\boldsymbol{y}^{*\prime}_{(t-1)s}\boldsymbol{M}_{ts}\boldsymbol{\varepsilon}^*_{ts}) = -\sigma^2_\varepsilon\sum_{s=S}^{T-1}\sum_{t=s+1}^{T}m_{ts}\alpha^{s-1}. \tag{4.45}$$

Note that derivations in (4.42)–(4.44) hold for both LD and MD-LD as well – only the bounds of the sums in equations (4.41) and (4.45) will differ (depending on the equations used).

To evaluate the biases of the estimators, we consider first the case with all possible instruments included: $m_{ts} = s$ for all $t > s$. For PD-LD, we obtain by Lemma 3 that

$$\begin{aligned}
\mathrm{E}(\boldsymbol{g}'_{nT^*}\boldsymbol{\varepsilon}^*) &= -\sigma^2_\varepsilon\sum_{s=S}^{T-1}\sum_{t=s+1}^{T}s\alpha^{s-1} = -\sigma^2_\varepsilon\sum_{s=S}^{T-1}(T-s)s\alpha^{s-1} \\
&= -\frac{\sigma^2_\varepsilon}{\alpha}\left[T\left(\frac{\alpha^S-\alpha^T}{(1-\alpha)^2} - \frac{(T-1)\alpha^T-(S-1)\alpha^S}{1-\alpha}\right)\right. \\
&\quad - 2\frac{\alpha^S-\alpha^T}{(1-\alpha)^3} + \frac{[2T-3]\alpha^T-[2S-3]\alpha^S}{(1-\alpha)^2} \\
&\quad \left. + \frac{(T-1)^2\alpha^T-(S-1)^2\alpha^S}{1-\alpha}\right],
\end{aligned} \tag{4.46}$$

which is of order $O(T^2\alpha^{T-\sqrt{2T}})$ for $T \to \infty$ as $S > T - \sqrt{2T}$. Similarly for MD-LD, it holds

$$\mathrm{E}(\boldsymbol{g}'_{nT^*}\boldsymbol{\varepsilon}^*) = -\sigma_\varepsilon^2 \sum_{s=2}^{T-1} \sum_{t=s+1}^{s+1} s\alpha^{s-1} = \left[-\frac{\sigma_\varepsilon^2}{\alpha} \left(\frac{\alpha^2 - \alpha^T}{(1-\alpha)^2} - \frac{(T-1)\alpha^T - \alpha^2}{1-\alpha} \right) \right], \tag{4.47}$$

which is of order $O(1)$ when $T \to \infty$. Finally, we have for LD

$$\mathrm{E}(\boldsymbol{g}'_{nT^*}\boldsymbol{\varepsilon}^*) = -\sigma_\varepsilon^2 \sum_{s=T-1}^{T-1} \sum_{t=T}^{T} s\alpha^{s-1} = -\sigma_\varepsilon^2 (T-1)\alpha^{T-2} = O(T\alpha^T). \tag{4.48}$$

Next, the case of a bounded number of instruments is considered: suppose $m_{ts} = \max(s, \bar{m})$ for all $t > s$. For PD-LD we have again by Lemma 3

$$\begin{aligned} |\,\mathrm{E}(\boldsymbol{g}'_{nT^*}\boldsymbol{\varepsilon}^*)| &= \left| -\sigma_\varepsilon^2 \sum_{s=S}^{T-1} \sum_{t=s+1}^{T} m_{ts}\alpha^{s-1} \right| \le \sigma_\varepsilon^2 \bar{m} \sum_{s=S}^{T-1} (T-s)|\alpha|^{s-1} \\ &= \frac{\sigma_\varepsilon^2 \bar{m}}{|\alpha|} \left(T\frac{|\alpha|^S - |\alpha|^T}{1-|\alpha|} + \frac{|\alpha|^S - |\alpha|^T}{(1-|\alpha|)^2} + \frac{|(S-1)|\alpha|^S - (T-1)|\alpha|^T|}{1-|\alpha|} \right), \end{aligned} \tag{4.49}$$

which is of order $O(T|\alpha|^{T-\sqrt{2T}})$ when $T \to \infty$. Similarly, it holds for MD-LD that

$$\begin{aligned} |\,\mathrm{E}(\boldsymbol{g}'_{nT^*}\boldsymbol{\varepsilon}^*)| &= \left| \sigma_\varepsilon^2 \sum_{s=2}^{T-1} \sum_{t=s+1}^{s+1} m_{ts}\alpha^{s-1} \right| \le \sigma_\varepsilon^2 \bar{m} \sum_{s=2}^{T-1} \sum_{t=s+1}^{s+1} |\alpha|^{s-1} \\ &= \frac{\sigma_\varepsilon^2 \bar{m}}{|\alpha|} \frac{1-|\alpha|^T}{|\alpha|^2 - |\alpha|}, \end{aligned} \tag{4.50}$$

which is of order $O(1)$ when $T \to \infty$. Finally, we can write for LD

$$\begin{aligned} |\,\mathrm{E}(\boldsymbol{g}'_{nT^*}\boldsymbol{\varepsilon}^*)| &= \left| \sigma_\varepsilon^2 \sum_{s=T-1}^{T-1} \sum_{t=T}^{T} m_{ts}\alpha^{s-1} \right| \le \sigma_\varepsilon^2 \bar{m} \sum_{s=T-1}^{T-1} \sum_{t=T}^{T} |\alpha|^{s-1} = \sigma_\varepsilon^2 \bar{m} |\alpha|^{T-2} \\ &= O(|\alpha|^T). \end{aligned} \tag{4.51}$$

□

4.A.2 Asymptotic distribution

The common notation will be discussed first. The proof of Theorem 3 is identical for LD, MD-LD, and PD-LD except for the dimensions of the instrument and data matrices used. Similarly to $\boldsymbol{y}^*$ in (4.16), let $\boldsymbol{W}^* = ([\boldsymbol{D}\boldsymbol{W}_1]', \ldots, [\boldsymbol{D}\boldsymbol{W}_n]')'$ and $\boldsymbol{W}_i^* = \boldsymbol{D}\boldsymbol{W}_i$, $i = 1, \ldots, n$, where $\boldsymbol{D}$ is the difference-operator matrix corresponding to the analyzed estimator. The instrument matrices $\boldsymbol{Z}$ and $\boldsymbol{Z}_i$ are also assumed to be corresponding to the estimator of interest (LD, MD-LD, or PD-LD). We will generically refer to $\hat{\boldsymbol{\theta}}_n$ as one of the estimator in this class, which can be now expressed as

$$\hat{\boldsymbol{\theta}}_n = \left(\boldsymbol{W}^{*\prime}\hat{\boldsymbol{Z}}\hat{\boldsymbol{V}}_n^{-1}\hat{\boldsymbol{Z}}'\boldsymbol{W}^*\right)^{-1}\boldsymbol{W}^{*\prime}\hat{\boldsymbol{Z}}\hat{\boldsymbol{V}}_n^{-1}\hat{\boldsymbol{Z}}'\boldsymbol{y}^*, \tag{4.52}$$

where $\hat{\boldsymbol{Z}}'\boldsymbol{W}^* = \sum_{i=1}^n \hat{\boldsymbol{Z}}_i'\boldsymbol{W}_i^*$, $\hat{\boldsymbol{Z}}'\boldsymbol{y}^* = \sum_{i=1}^n \hat{\boldsymbol{Z}}_i'\boldsymbol{y}_i^*$, and the instrument matrix $\hat{\boldsymbol{Z}}_i$ refers to the feasible counterpart of $\boldsymbol{Z}_i$. Given that the $T^* \times R$ matrix $\hat{\boldsymbol{Z}}_i$ is block diagonal, $\hat{\boldsymbol{Z}}_i = \operatorname{diag}(\hat{\boldsymbol{z}}_{its}')$, the $R \times (K+1)$ matrix $\hat{\boldsymbol{Z}}_i'\boldsymbol{W}_i^*$ can be conveniently partitioned in vectors in the following way

$$\hat{\boldsymbol{Z}}_i'\boldsymbol{W}_i^* = (\hat{\boldsymbol{z}}_{its}w_{itsk}^*)_{(t,s)\in\mathcal{T},k=1,\ldots,K+1}, \tag{4.53}$$

where $(t,s) \in \mathcal{T}$ is the running row-index with values depending on the type of estimator,

$$\begin{aligned}
\mathcal{T}_{\mathrm{LD}} &= \left\{(t,s) : t = T;\ s = T-1\right\}, \\
\mathcal{T}_{\mathrm{MD\text{-}LD}} &= \left\{(t,s) : t = s+1;\ s = 2, \ldots, T-1\right\}, \\
\mathcal{T}_{\mathrm{PD\text{-}LD}} &= \left\{(t,s) : t = s+1, \ldots, T;\ s = S, \ldots, T-1\right\},
\end{aligned} \tag{4.54}$$

and $k = 1, \ldots, K+1$ is the column index.

The following lemmas will now analyze individual terms of

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^0) = \left(\frac{\boldsymbol{W}^{*\prime}\hat{\boldsymbol{Z}}}{n}\hat{\boldsymbol{V}}_n^{-1}\frac{\hat{\boldsymbol{Z}}'\boldsymbol{W}^*}{n}\right)^{-1}\frac{\boldsymbol{W}^{*\prime}\hat{\boldsymbol{Z}}}{n}\hat{\boldsymbol{V}}_n^{-1}\frac{\hat{\boldsymbol{Z}}'\boldsymbol{\varepsilon}^*}{\sqrt{n}}, \tag{4.55}$$

which is obtained by substituting for $\boldsymbol{y}^*$ in (4.52) from model (4.25) and where the notation $\boldsymbol{\varepsilon}^* = ([\boldsymbol{D}\boldsymbol{\varepsilon}_1]', \ldots, [\boldsymbol{D}\boldsymbol{\varepsilon}_n]')'$ is used.

Lemma 4. *Suppose Assumptions A.1–A.5 hold for a fixed T and $n \to \infty$. Then*

$$\frac{1}{n}\hat{\boldsymbol{Z}}'\boldsymbol{W}^* \xrightarrow{p} \boldsymbol{\Omega}. \tag{4.56}$$

Proof. We use the decomposition

$$\frac{1}{n}\hat{\boldsymbol{Z}}'\boldsymbol{W}^* = \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{Z}}_i'\boldsymbol{W}_i^* = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\boldsymbol{z}}_{its}w_{itsk}^*\right)_{(t,s)\in\mathcal{T};k=1,\ldots,K+1}, \tag{4.57}$$

where (t,s) and k are the row and column indices, respectively, of the matrix $\hat{\boldsymbol{Z}}_i\boldsymbol{W}_i^*$. Next, let us analyze the generic vector

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{z}}_{its}w_{itsk}^* &= \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{z}_{its} - \boldsymbol{\mathcal{W}}_{its}(\hat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}^0)\right)w_{itsk}^* \\ &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_{its}w_{itsk}^* - \left[\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\mathcal{W}}_{its}w_{itsk}^*)\right](\hat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}^0).\end{aligned} \tag{4.58}$$

First, note that $\sum_{i=1}^{n}\boldsymbol{z}_{its}w_{itsk}^*/n \to \boldsymbol{\omega}_{tsk} = \mathrm{E}(\boldsymbol{z}_{its}w_{itsk}^*)$ in probability as $n \to \infty$ by the law of large numbers (Davidson, 1994, Theorem 20.8) and Assumptions A.1 and A.5. The same argument applies to $\sum_{i=1}^{n}(\boldsymbol{\mathcal{W}}_{its}w_{itsk}^*)/n$. Finally, $\hat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}^0 = o_p(1)$ follows from the consistency of the preliminary estimator $\hat{\boldsymbol{\theta}}_n^0$ (see Assumption A.3). Consequently, $\sum_{i=1}^{n}\hat{\boldsymbol{z}}_{its}w_{itsk}^*/n \to \boldsymbol{\omega}_{tsk}$ in probability as $n \to \infty$ for any t, s, and k and we can rewrite (4.57) as

$$\frac{1}{n}\hat{\boldsymbol{Z}}'\boldsymbol{W}^* = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\boldsymbol{z}}_{its}w_{itsk}^*\right)_{ts,k} = \boldsymbol{\Omega} + o_p(1). \tag{4.59}$$

□

Lemma 5. *Suppose Assumptions A.1–A.5 hold for a fixed T and $n \to \infty$. Then*

$$\frac{1}{\sqrt{n}}\hat{\boldsymbol{Z}}'\boldsymbol{\varepsilon}^* \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{M\Sigma M}'). \tag{4.60}$$

Proof. We use again the decomposition

$$\frac{1}{\sqrt{n}}\hat{\boldsymbol{Z}}'\boldsymbol{\varepsilon}^* = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{Z}}_i'\boldsymbol{\varepsilon}_i^* = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\hat{\boldsymbol{z}}_{its}\varepsilon_{its}^*\right)_{(t,s)\in\mathcal{T}}. \tag{4.61}$$

Next, let us analyze the generic vector and substitute for the initial estimator from (4.26):

$$
\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{z}}_{its}\varepsilon^*_{its} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\boldsymbol{z}_{its} - \frac{1}{\sqrt{n}}\boldsymbol{\mathcal{W}}_{its}\sqrt{n}(\hat{\boldsymbol{\theta}}^0_n - \boldsymbol{\theta}^0)\right)\varepsilon^*_{its} \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{z}_{its}\varepsilon^*_{its} - \frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{\mathcal{W}}_{its}\left(\boldsymbol{\Lambda}\sqrt{n}\boldsymbol{f}_n(\boldsymbol{\theta}^0) + o_p(1)\right)\varepsilon^*_{its}\right] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{z}_{its}\varepsilon^*_{its} - \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\mathcal{W}}_{its}\varepsilon^*_{its}\right)\boldsymbol{\Lambda}\sqrt{n}\boldsymbol{f}_n(\boldsymbol{\theta}^0) + o_p(1).
\end{aligned}
\tag{4.62}
$$

The law of large numbers (Davidson, 1994, Theorem 20.8) and Assumptions A.1 and A.5 imply that $\sum_{i=1}^{n}\boldsymbol{\mathcal{W}}_{its}\varepsilon^*_{its}/n \to \boldsymbol{P}_{ts} = \mathrm{E}(\boldsymbol{\mathcal{W}}_{its}\varepsilon^*_{its})$ for each t and s.

As $\boldsymbol{P} = (\boldsymbol{P}'_{(S+1)S}, \ldots, \boldsymbol{P}'_{T(T-1)})'$, we can rewrite (4.61) as

$$
\begin{aligned}
\frac{1}{\sqrt{n}}\hat{\boldsymbol{Z}}'\boldsymbol{\varepsilon}^* &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\hat{\boldsymbol{z}}_{its}\varepsilon^*_{its}\right)_{(t,s)\in\mathcal{T}} \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\boldsymbol{z}_{its}\varepsilon^*_{its})_{(t,s)\in\mathcal{T}} - (\boldsymbol{P}_{ts})_{(t,s)\in\mathcal{T}}\boldsymbol{\Lambda}\sqrt{n}\boldsymbol{f}_n(\boldsymbol{\theta}^0) + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{Z}_i'\boldsymbol{\varepsilon}_i^* - \boldsymbol{P\Lambda}\sqrt{n}\boldsymbol{f}_n(\boldsymbol{\theta}^0) + o_p(1) \qquad (4.63) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\tau}_i(\boldsymbol{\theta}^0) - \boldsymbol{P\Lambda}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\psi}_i(\boldsymbol{\theta}^0) + o_p(1) \\
&= \boldsymbol{M}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\rho}_i(\boldsymbol{\theta}^0) + o_p(1),
\end{aligned}
$$

where $\boldsymbol{M} = (\boldsymbol{I}_R, -\boldsymbol{P\Lambda})$ and $\boldsymbol{\tau}_i(\boldsymbol{\theta}^0) = \boldsymbol{Z}_i'\boldsymbol{\varepsilon}_i^* = \boldsymbol{Z}_i'\boldsymbol{D}\boldsymbol{\varepsilon}_i$ denotes the moment conditions of the LD-type estimator at $\boldsymbol{\theta}^0$.

By Assumption A.1 and A.4, $\boldsymbol{\rho}_i(\boldsymbol{\theta}^0)$ are independent random vectors satisfying $\mathrm{E}[\boldsymbol{\rho}_i(\boldsymbol{\theta}^0)] = \boldsymbol{0}$. As the second and higher moments exist by Assumptions A.4, the central limit theorem (Davidson, 1994, Theorem 23.12 and 25.6) imply

$$
\frac{1}{\sqrt{n}}\hat{\boldsymbol{Z}}'\boldsymbol{\varepsilon}^* = \boldsymbol{M}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\rho}_i(\boldsymbol{\theta}^0) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{M\Sigma M}'). \tag{4.64}
$$

□

Proof of Theorem 3. Let $\hat{\boldsymbol{\theta}}_n$ be either LD, PD-LD or MD-LD in (4.52). By

(4.55), it can be written as

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) = \left[\left(\frac{\boldsymbol{W}^{*\prime}\hat{\boldsymbol{Z}}}{n}\right)\hat{\boldsymbol{V}}_n^{-1}\left(\frac{\hat{\boldsymbol{Z}}^{\prime}\boldsymbol{W}^{*}}{n}\right)\right]^{-1}\left(\frac{\boldsymbol{W}^{*\prime}\hat{\boldsymbol{Z}}}{n}\right)\hat{\boldsymbol{V}}_n^{-1}\left(\frac{\hat{\boldsymbol{Z}}^{\prime}\boldsymbol{\varepsilon}^{*}}{\sqrt{n}}\right).$$

By Assumption A.4 and Lemma 4, it follows for $n \to \infty$

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) = \left(\left[\boldsymbol{\Omega}^{\prime}\boldsymbol{V}\boldsymbol{\Omega}\right]^{-1}\boldsymbol{\Omega}^{\prime}\boldsymbol{V} + o_p(1)\right)\left(\frac{\hat{\boldsymbol{Z}}^{\prime}\boldsymbol{\varepsilon}^{*}}{\sqrt{n}}\right).$$

The claim of the theorem now follows from Lemma 5. □

Bibliography

Abrevaya, J. (2000, 3). Rank estimation of a generalized fixed-effects regression model. *Journal of Econometrics 95*(1), 1–23.

Ahn, S. C. and P. Schmidt (1997, 1). Efficient estimation of dynamic panel data models: Alternative assumptions and simplified estimation. *Journal of Econometrics 76*(1-2), 309–321.

Anderson, T. W. and C. Hsiao (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics 18*(1), 47–82.

Aquaro, M. and P. Čížek (2010). One-step robust estimation of fixed-effects panel data models. Technical Report 2010-110, Tilburg University, Center for Economic Research.

Aquaro, M. and P. Čížek (2013, 1). One-step robust estimation of fixed-effects panel data models. *Computational Statistics & Data Analysis 57*(1), 536–548.

Arellano, M. (2003). *Panel Data Econometrics*. Number 9780199245291 in OUP Catalogue. Oxford University Press.

Arellano, M. and S. Bond (1991, 4). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The Review of Economic Studies 58*(2), 277–297.

Arellano, M. and O. Bover (1995, 7). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics 68*(1), 29–51.

Azzalini, A. and M. G. Genton (2008). Robust likelihood methods based on the skew-t and related distributions. *International Statistical Review 76*(1), 106–129.

Bain, L. J. and M. Engelhardt (1992). *Introduction to probability and mathematical statistics*. Pacific Grove, CA: Duxbury.

Baltagi, B. H. (2005). *Econometric analysis of panel data*. John Wiley & Sons.

Baltagi, B. H. and G. Bresson (2011). A robust Hausman-Taylor estimator. Unpublished manuscript.

Blundell, R. and S. Bond (1998, 8). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics 87*(1), 115–143.

Bramati, M. C. and C. Croux (2007). Robust estimators for the fixed effects panel data model. *Econometrics Journal 10*(3), 521–540.

Bun, M. J. and M. A. Carree (2005). Bias-corrected estimation in dynamic panel data models. *Journal of Business & Economic Statistics 23*, 200–210.

Bun, M. J. and J. F. Kiviet (2006, 6). The effects of dynamic feedbacks on ls and mm estimator accuracy in panel data models. *Journal of Econometrics 132*(2), 409–444.

Carrasco, M. and J.-P. Florens (2000). Generalization of gmm to a continuum of moment conditions. *Econometric Theory 16*(06), 797–834.

Čížek, P. (2006, November). Least trimmed squares in nonlinear regression under dependence. *Journal of Statistical Planning and Inference 136*(11), 3967–3988.

Čížek, P. (2008, June). Robust and efficient adaptive estimation of binary-choice regression models. *Journal of the American Statistical Association 103*, 687–696.

Čížek, P. (2010, November). Reweighted least trimmed squares: an alternative to one-step estimators. Discussion Paper 91/2010. CentER, Tilburg University.

Čížek, P. (2011, 01). Semiparametrically weighted robust estimation of regression models. *Computational Statistics & Data Analysis 55*(1), 774–788.

Čížek, P. (2012, 6). Semiparametric robust estimation of truncated and censored regression models. *Journal of Econometrics 168*(2), 347–366.

Coakley, J., A.-M. Fuertes, and R. Smith (2006, 05). Unobserved heterogeneity in panel time series models. *Computational Statistics & Data Analysis 50*(9), 2361–2380.

Cornwell, C. and P. Rupert (1988). Efficient estimation with panel data: An empirical comparison of instrumental variables estimators. *J. Appl. Econ. 3*(2), 149–155.

Davidson, J. (1994). *Stochastic Limit Theory.* New York: Oxford University Press.

Davies, P. L. and U. Gather (2005). Breakdown and groups. *The Annals of Statistics 33*(3), 977–1035.

Dhaene, G. and Y. Zhu (2009). Median-based estimation of dynamic panel models with fixed effects. Unpublished manuscript.

Galvao, Jr., A. F. (2011). Quantile regression for dynamic panel data with fixed effects. *Journal of Econometrics 164*(1), 142–157.

Genton, M. G. and A. Lucas (2003). Comprehensive definitions of breakdown points for independent and dependent observations. *Journal of the Royal Statistical Society Series B 65*(1), 81–94.

Gervini, D. and V. J. Yohai (2002, 4). A class of robust and fully efficient regression estimators. *The Annals of Statistics 30*(2), 583–616.

Gouriéroux, C., P. C. Phillips, and J. Yu (2010). Indirect inference for dynamic panel models. *Journal of Econometrics 157*(1), 68–77.

Griliches, Z. and J. A. Hausman (1986). Errors in variables in panel data. *Journal of Econometrics 31*(1), 93 – 118.

Hahn, J. (1999, 12). How informative is the initial condition in the dynamic panel model with fixed effects? *Journal of Econometrics 93*(2), 309–326.

Hahn, J., J. Hausman, and G. Kuersteiner (2007, 10). Long difference instrumental variables estimation for dynamic panel models with fixed effects. *Journal of Econometrics 140*(2), 574–617.

Hahn, J. and G. Kuersteiner (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and t are large. *Econometrica 70*(4), 1639–1657.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986). *Robust statistics: the approach based on influence functions*. John Wiley & Sons.

Harris, M. N., L. Mátyás, and P. Sevestre (2008). Dynamic models for short panels. In L. Mátyás and P. Sevestre (Eds.), *The Econometrics of Panel Data*, Chapter 8, pp. 249–278. Springer-Verlag.

Hayakawa, K. (2009, 12). On the effect of mean-nonstationarity in dynamic panel data models. *Journal of Econometrics 153*(2), 133–135.

Holtz-Eakin, D., W. Newey, and H. S. Rosen (1988, 11). Estimating vector autoregressions with panel data. *Econometrica 56*(6), 1371–1395.

Honoré, B. E. and J. L. Powell (2005). Pairwise difference estimation of nonlinear models. In D. W. K. Andrews and J. H. Stock (Eds.), *Identification and Inference for Econometric Models. Essays in Honor of Thomas Rothenberg.*, Chapter 22, pp. 520–553. Cambridge University Press.

Kiviet, J. F. (1995, 7). On bias, inconsistency, and efficiency of various estimators in dynamic panel data models. *Journal of Econometrics 68*(1), 53–78.

Kiviet, J. F. (2007). Judging contending estimators by simulation: tournaments in dynamic panel data models. In G. Phillips and E. Tzavalis (Eds.), *The Refinement of Econometric Estimation and Test Procedures*, Chapter 11, pp. 282–318. Cambridge University Press.

Koenker, R. (2004, 10). Quantile regression for longitudinal data. *Journal of Multivariate Analysis 91*(1), 74–89.

Lucas, A., R. van Dijk, and T. Kloek (2007). Outlier robust GMM estimation of leverage determinants in linear dynamic panel data models. Unpublished manuscript.

Maronna, R. A., R. D. Martin, and V. J. Yohai (2006). *Robust Statistics*. John Wiley & Sons.

Maronna, R. A. and V. J. Yohai (2000, 8). Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference 89*(1-2), 197–214.

Mátyás, L. and P. Sevestre (Eds.) (2008). *The Econometrics of Panel Data*, Volume 46. Springer.

Muler, N., D. Peña, and V. J. Yohai (2009, April). Robust estimation for arma models. *The Annals of Statistics 37*(2), 816–840.

Nerlove, M. (2005). *Essays in Panel Data Econometrics.* Number 9780521022460. Cambridge University Press.

Newey, W. K. and R. J. Smith (2004, 1). Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica 72*(1), 219–255.

Nickell, S. (Nov., 1981). Biases in dynamic models with fixed effects. *Econometrica 49*(6), 1417–1426.

Ronchetti, E. and F. Trojani (2001, 3). Robust inference with gmm estimators. *Journal of Econometrics 101*(1), 37–69.

Rousseeuw, P. J. (1984, 12). Least median of squares regression. *Journal of the American Statistical Association 79*(388), 871–880.

Rousseeuw, P. J. and A. M. Leroy (1987). *Robust regression and outlier detection.* Wiley.

Simpson, D. G., D. Ruppert, and R. J. Carroll (1992, June). On one-step gm estimates and stability of inferences in linear regression. *Journal of the American Statistical Association 87*(418), 439–450.

Stromberg, A. J., O. Hossjer, and D. M. Hawkins (2000). The least trimmed differences regression estimator and alternatives. *Journal of the American Statistical Association 95*(451), 853–864.

Vandev, D. L. and N. M. Neykov (1998, 01). About regression estimators with high breakdown point. *Statistics 32*(2), 111–129.

Wagenvoort, R. and R. Waldmann (2002, February). On b-robust instrumental variable estimation of the linear model with panel data. *Journal of Econometrics 106*(2), 297–324.

Wooldridge, J. M. (2001). *Econometric Analysis of Cross Section and Panel Data*, Volume 1 of *MIT Press Books.* The MIT Press.

Zaman, A., P. J. Rousseeuw, and M. Orhan (2001, 4). Econometric applications of high-breakdown robust regression techniques. *Economics Letters 71*(1), 1–8.